The Dissertation Committee for Hui Wu Certifies that this is the approved version of the following dissertation:

A FRAMEWORK FOR DEVELOPING ROAD RISK INDICES USING QUANTILE REGRESSION BASED CRASH PREDICTION MODEL

Committee:

Zhanmin Zhang, Supervisor

Michael Murphy

Randy B. Machemehl

Steven T. Waller

Elmira Popova

A FRAMEWORK FOR DEVELOPING ROAD RISK INDICES USING QUANTILE REGRESSION BASED CRASH PREDICTION MODEL

by

Hui Wu, B.E., M.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2011

To my parents and husband for their unconditional love and support, to my daughter for always giving me happiness

Acknowledgements

My graduate studies have been benefitted greatly from the assistance and guidance of a number of people. First and foremost, I must express my sincere appreciation to my supervisor Dr. Zhanmin Zhang his invaluable guidance, illuminating ideas and suggestions, understanding and endless encouragement through my Ph.D. studies. His technical and editorial advice was essential for this research. It has been a very fulfilling and rewarding experience.

My sincere thanks also go to the members of my dissertation committee, namely Dr. Michael Murphy from the Center of Transportation Research at UT, Dr. Randy B. Machemehl and Dr. Steven T. Waller from the Civil, Architectural and Environmental Engineering Department, and Dr. Elmira Popova from the Department of Mechanical Engineering. Your help and support was very significant and your feedback indispensable to the development of my dissertation and its successful completion.

I want to express my gratitude towards my parents for all the love and support during my graduate studies and throughout my life. Thank you for supporting both emotionally and financially my academic adventures and for being there for me in times of hardship and doubt. I would also like to express my deep gratitude to my husband, Qi Li, for his patience, understanding, love and encouragement. Last but not least,, I am grateful to the all the friends and current and former colleagues at UT Aristeidis Pantelias, Lu Gao, Sunny Jaipuria, Sruthi Sree Sravya Peddibhotla, Wenxing Liu, Liang Liang, Abdus Qazi, Jea Won Hwang, Weiyuan Yuwen, Epi Gonzalez, Jin Liu, Rou Luo, Runhua Guo, Jianming Ma, and Brenda B. Zhou, for their friendship and kind help. I would also like to show my special gratitude to Mandy D. Weyant for her never failing willingness to assist and encourage.

A FRAMEWORK FOR DEVELOPING ROAD RISK INDICES USING QUANTILE REGRESSION BASED CRASH PREDICTION MODEL

Publication No._____

Hui Wu, Ph.D.

The University of Texas at Austin, 2011

Supervisor: Zhanmin Zhang

Safety reviews of existing roads are becoming a popular practice of many agencies nationally and internationally. Knowing road safety information is of great importance to both policymakers in addressing safety concerns and travelers in managing their trips. There have been various efforts in developing methodologies to measure and assess road safety in an effective manner. However, the existing research and practices are still constrained by their subjective and reactive nature.

The goal of this research is to develop a framework of Road Risk Indices (RRIs) to assess road risks of existing highway infrastructure for both road users and agencies based on road geometrics, traffic conditions, and historical crash data. The proposed RRIs are intended to give a comprehensive and objective view of road safety, so that safety problems can be identified at an early stage before they rise in the form of

accidents. A methodological framework of formulating RRIs that integrates results from crash prediction models and historical crash data is proposed, and Linear Referencing tools in the ArcGIS software are used to develop digital maps to publish estimated RRIs. These maps provide basic Geographic Information System (GIS) functions, including viewing and querying RRIs, and performing spatial analysis tasks. A semi-parameter count model and quantile regression based estimation are proposed to capture the specific characteristics of crash data and provide more robust and accurate predictions on crash counts. Crash data collected on Interstate Highways in Washington State for the year 2002 was extracted from the Highway Safety Information System (HSIS) and used for the case study. The results from the case study show that the proposed framework is capable of capturing statistical correlations between traffic crashes and influencing factors, leading to the effective integration of safety information in composite indices.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 INTRODUCTION

1.1 Background and Motivation

Despite notable improvements in roadway design, vehicle safety, and operational decisions, traffic crashes still remain at a relatively high level, posing a major public health and injury prevention problem. According to the World Health Organization, more than a million people are killed on the world's roads each year (WHO, 2010). In the U.S., more than 37,000 people died in road traffic crashes every year within the past 10 years, and the property damage caused by traffic crashes were estimated at more than 200 billion each year (NHTSA, 2009).

In practice, measures derived from historical traffic crash counts (i.e., crash frequency, crash rate, Equivalent Property Damage Only average crash frequency, etc.) are widely used to assess safety performance of existing highway infrastructures (ASSHTO, 2011). These types of measures are simple, but they are imperfect indicators for several reasons: 1) these measures assess road safety in a reactive manner by using historical data, and thus they cannot proactively prevent the occurrence of crashes; 2) historical traffic crash counts are not always available, and obtaining crash data is unlikely for newly built highways; 3) the number of traffic crashes is subject to random fluctuations, particularly in the sense that short-term patterns do not necessarily reflect long-term trends (Hakkert et al, 2007); 4) these measures do not provide enough information for researchers to understand the processes that contribute to crashes and subsequently identify ways to improve road safety.

For the selected high crash locations, road safety assessment program is carried out. Road safety assessment aims to identify potential hazards by measuring risk in relation to road features, so that remedial treatments may be implemented to reduce the probability of or prevent future crashes. Current road safety assessment in the U.S. is implemented through the Road Safety Audits (RSAs) program under the Federal Highway Administration (FHWA). The program intent is to reduce high crash occurrence locations on the road by making possible modifications to existing roads or improving the design of similar new roads. RSAs have been used widely in the United States since 1997 and are also commonly used in the United Kingdom and Australia. In the program, an independent multidisciplinary team of professionals with varied expertise is usually formed to qualitatively estimate and report potential road safety issues and seek ways to improve safety for all road users (FWHA, 2009). Typical forms of the reports include three components: identified problems, potential causes, and engineer suggestions.

Serving as a valuable tool that gives a view of safety issues with support from safety experts, the process is subjective, heavily depending on the individuals' experience and judgment. Also, because the assessment is case-by-case based, applying the results to decision-making about safety improvement activities at the network level is difficult. A well-defined methodology for systematically quantifying safety performance of existing roads is therefore needed. Such a methodology should provide transportation professionals with tools that help them clearly consider safety when making decisions related to the management of transportation facilities, notably highways, in an objective and proactive manner. In addition, the tool should be able to provide road users with

comprehensive information about the safety conditions of the highway facilities and in turn help them in the route selection process.

In order to address the issues discussed above, this research is focused on a methodology of developing composite Road Risk Indices (RRIs) to quantify the potential road risks. While an indicator is a quantitative or a qualitative measure derived from a series of observed facts that can reveal relative positions in an area (Nardo et al., 2005), composite indices aggregate various kinds of information in the indicators.

It should be emphasized that these indices are intended to show statistical correlations between roadway characteristics, traffic conditions, and crashes, but not necessarily to represent cause-and-effect relationships. Thus, RRIs should be applied to reflect relative safety levels of the facilities for the purpose of safety management, rather than serve as explanations for a particular accident.

As in other risk evaluation methodologies, RRIs focus on crash frequencies and their related loss. Statistical models that describe the relation between crash frequencies and their influencing factors have been widely studied for the last few decades. Results from crash prediction models contain a significant amount of useful information and are mainly used in the design phase. In this research, a quantile regression based crash prediction model is proposed to develop RRIs and assess safety for the existing roads.

The number of traffic crashes is a random variable with its distinct characteristics: discrete and non-negative integer with a large proportion of zeros, and the remaining values being skewed toward the right. Because of these characteristics, techniques for modeling count data are commonly used to analyze crash frequency and its influencing

factors such as geometrics, traffic conditions, climate, road user attributes, and vehicular-related characteristics. To describe the conditional distribution of a count outcome, two categories of count models can be used: fully parametric probabilistic models and non-or semiparametric probabilistic models.

Fully parametric probabilistic models, including Poisson, negative binomial (NB) regression models and their variants, have been extensively used in the studies of crash frequency over past decades. Despite all the advantages of fully parametric probabilistic count data models in terms of modeling, estimation, and inference, there are certain drawbacks preventing reliance on these models as the predictive analysis device. Fully parametric probabilistic models impose restrictive parametric assumptions in the way that the covariates affect the response variable. As a consequence, those models usually lack robustness, even when flexible models like mixed models are applied. Moreover, fully parametric models could result in inconsistent or very poor estimation and inference under inappropriate distribution assumptions. The limitation of their ability to handle heterogeneity is another big concern. Given these limitations, it is attractive to study how non or semiparametric models which freely approximate the conditional distribution perform for crash modeling.

Quantile regression (QR) for counts is one of the semiparametric models that can be used as a methodological alternative in analyzing crash frequency. Compared with existing models, the proposed model provides a more complete and robust analysis of crash data for at least two reasons. First, crash data usually follows typical count distributions with a large proportion of zeros and the remaining values highly skew

toward the right. Quantile regression becomes appealing in terms of providing a more complete picture of effects of covariates on crash frequency rather than just the mean because it estimates various quantiles of a population. Second, as a semiparametric model, quantile regression for counts allows researchers to relax restrictions in the form of the distribution function of the response variable, resulting in more robust estimation.

1.2 Research Goal and Objectives

The goal of this research is to develop a framework of RRIs to assess road risks of existing highway infrastructure for both road users and agencies based on road geometrics, traffic conditions, and historical crash data. The proposed framework should be capable of identifying statistical correlations between traffic crashes and influencing factors and integrating useful information in composite indices. To achieve this goal, the following objectives are expected to be accomplished:

1. Formulate a framework to develop RRIs that provide a basis for assessing road risks of existing highway infrastructure for both road users and agencies. The proposed RRIs should be able to identify the nature of road risk and reflect relative safety levels of the facilities for the purpose of safety management in an objective and proactive manner. Also, the framework should be easy to adopt state-of-the-art crash prediction models and flexible to accommodate the historical data availability.

2. Develop crash models for predicting traffic crash frequency using quantile regression for count data. Such models should be able to investigate the "complete picture" rather than just the mean of the relation between explanatory variables and crash frequency, and further give reliable and robust predictions. In addition, the proposed crash prediction model should be compatible with the proposed RRIs so that the results can be easily integrated with the development of RRIs.
3. Develop a geographic information system (GIS) based platform to visualize the developed RRIs. Such a system should be able to offer tools for viewing and querying the spatial and attribute data.

1.3 Research Scope

The scope of this research is to develop a framework that is capable of assessing road risks of existing highway infrastructure for both road users and agencies based on road geometrics, traffic conditions, and historical crash data. To demonstrate the methodology, a dataset containing crash data and corresponding road inventory data extracted from Highway Safety Information System (HSIS) is employed to conduct numerical analysis for model calibration and validation. The proposed research focuses on developing RRIs for highway segments, but the framework including the techniques in building the QR based crash prediction model can be directly adopted in developing RRIs for intersections.

1.4 Contributions

This research will benefit both road users and decision makers through the provision of indices that assess road safety performance in an objective and proactive manner. Contributions of this research include:

1. The development of a methodological framework that can be used to assess road safety of existing roads through composite indices, taking into account the roadway characteristics, traffic conditions and historical crash data,
2. The development of a detailed formulation of RRIs that is capable for integrating the results of crash prediction models and historical crash data, and providing road safety information to both road users and decision makers in an effective way,
3. The application of quantile regression techniques in traffic crash modeling, which provides a more robust and fuller analyses for crash data, and
4. The development of GIS maps as a user-friendly interface for publishing RRIs.

1.5 Dissertation Outline

This chapter briefly introduces the concepts of using indicators as a proactive and objective tool to assess road risk, as well as the goals and contributions of the dissertation. The remainder of the dissertation is arranged as follows:

Chapter 2 presents a comprehensive literature review of the various topics that form the background of this research. These topics include state of the art and the practice in road safety management, research on traffic crash occurrence modeling, and Quantile Regression and its application to count data.

Chapter 3 describes the research methodology, including a framework used for outlining the development of the methodological framework and detailed discussions on the design and implementation of each parts of the proposed methodological framework. The discussion covers but not limited to the following topics: the formulation of the Road Risk Indices (RRIs), the application of quantile regression (QR) on counts techniques to the crash occurrence modeling, prediction methods using estimated conditional distribution from the QR based crash prediction model, the selection of the dataset for the empirical study, GIS techniques and tools to present the results.

In Chapter 4 and Chapter 5, a case study is presented to illustrate the implementation of the proposed methodology.

In order to demonstrate the application of the proposed QR based crash prediction model with real data, Chapter 4 presents a case study by applying the proposed model to the Interstate Highway (IH) system in Washington State. More specifically, crash data collected on Interstate Highways in Washington State for 2002 were extracted from Highway Safety Information System (HSIS) and used for model calibration and validation.

Chapter 5 discusses the process of developing RRIs using the estimated results from the previous chapter on a selected set of IH segments based on the road geometry

features, traffic characteristics, and historical crash data. Both the Probability Method and the Location Method are applied to calculate RRIs. The estimated RRIs are then visualized on digital maps using the linear referencing tools in ArcGIS. The developed maps support all the basic view and query functions of a typical Geographic Information System (GIS).

Chapter 6 summarizes the research effort and presents the conclusions, and identifies the directions for future research.

CHAPTER 2 LITERATURE REVIEW

This chapter presents an overview of the background literature in three major areas pertaining to this research: research and practice of developing road risk indices, traffic crash occurrence modeling methodologies, and quantile regression and its application to count data. In the first section, state-of-the-art studies on formulating Road Risk Indices (RRIs) for the purpose of assessing safety performance on existing roads are reviewed. In the second section, a thorough review of probabilistic traffic crash prediction models is presented. In the third section, a general background on quantile regression (QR) and the recent emerging techniques for QR on counts are introduced.

2.1 Road Risk Indices

Several studies have been carried out to develop indices for highway safety management purposes. De Leur and Sayed (2002) presented a methodology where a Road Safety Risk Index (RSRI) was developed as a driver-based subjective assessment of road risks of existing roads. Using data collected on the Trans Canada Highway (TCH) in British Columbia, the RSRI was compared with objectively derived road safety measures, and the results showed that they were statistically compatible. One of the limitations of the RSRI pointed out by the authors was the subjective nature of the process which might lead to doubts on accuracy, reliability and repeatability of the results.

Montella (2005) developed a Potential For safety Improvement (PFI) index to quantify the safety gains that could be achieved by addressing the problems identified in

the review process. The PFI is formulated by a factoring approach based on known crash relationships, and weights of each safety issues in the formulation are derived from existing literature. Crash data on rural two-lane highways at non-intersections in Italy was used for the model validation. Because cited studies can have different and sometimes conflicting assumptions, and the calibration processes are usually carried out using a different dataset, directly combining the contributions of road features to the road safety using a factoring method such as PFI may lead to inconsistent and unreliable estimates.

In practice, assessing safety performance of existing roads by composite indicators is becoming an accepted practice at the international level, especially in European countries, Canada, and New Zealand.

Launched by the Automobile Association (AA) Foundation for Road Safety Research in 2002, the European Road Assessment Programme (EuroRAP) aims to provide independent, consistent safety ratings of roads across borders. The program has grown from a 4-country pilot to a major force for change in over 20 members over the past few years (EuroRAP, 2006a).

One of the important contributions of this program is the development of a procedure for a driver to inspect routes and assess the Road Protection Score (RPS) (EuroRAP, 2006b). The RPS which ranges from 1 to 4, rates the safety of a road based on how well its design would protect a car user from being severely injured or killed if a head-on, run-off, or intersection crash occurs. The higher the scores, the better the highway design in terms of minimizing the severity of a collision. The RPSs are

evaluated based on field survey, in which a direct visual inspection of the road quality is performed and risk tables are developed based on speed limits and road design features by engineers. Compared with other road safety reviews, the RPS assesses the safety performance of a route rather than the homogeneous section or individual site of concern. Although being one of the best road safety performance indices available, the RPS has its weak side. The evaluation process totally relies on inspectors' judgment and thus is subjective and not reliable in all cases. Large data has been collected but further analysis is needed.

Risk mapping is another tool developed by EuroRAP to show the risk a driver faces on a certain road derived from the crash history of the road (EuroRAP, 2006c). On the map, traffic crash rates derived from historical crash data are displayed in risk rate bandings (low, low-medium, medium, medium-high, high). Compared to RPS, risk mapping provides an objective rating of the safety performance of existing roads, but it has all the shortcomings as other ratings simply based on crash data. First, the number of road crashes is subject to random fluctuations, where "short term change in the recorded numbers does not necessarily reflect a change in the underlying, long-term expected numbers."(Hakkert et al, 2007) Also, this rating system depends largely on the availability, integrity and quality of the historical crash records. Last but not least, it evaluates road safety in reactive manner rather than proactive manner.

In New Zealand, Transfund (now Land Transport New Zealand) started Road Infrastructure Safety Assessment (RISA) program in 2002, as an evidential based system for assessing the impact of the road engineering features on road safety (Transfund,

2003). A two-stage based rating methodology was developed to measure road safety performance of several sample networks. In the first stage, a team of professional assessors and a driver is sent out to collect all the necessary data and estimate relative risks for selected features, comparing the safety performance of a road section against a reference road section. In the second stage, an overall measure of road infrastructure safety provision is developed through adjusting the results from the first stage by factoring for road type, terrain type and traffic volume. In the RISA program, two independent teams are formed to assess the same sections to carry out the repeatability trial, and the results showed that the variation in assessing the Risk Level Ratings raised concerns about lack of repeatability (Transfund, 2003).

2.2 Traffic Crash Modeling

Statistical modeling of traffic crashes forms the basis for investigating and analyzing accidents, preventing future ones, and reducing losses in terms of personal injury and property damage. These models describe the relation between crash frequency and their influencing factors such as roadway geometrics, traffic conditions, highway user attributes, and vehicular- and crash-related characteristics. They can be used to improve road safety by identifying black spots, predicting motor vehicle crashes and developing road risk indices.

Extensive research on road crash prediction models has been performed for the last few decades. Considering the nature of the number of traffic accidents, which is usually discrete and non-negative integer with a large proportion of zeros and the

remaining values being skewed toward the right, techniques for modeling count data are commonly employed in modeling process. More specifically, two categories of count models can be used to describe the conditional distribution of a count outcome: fully parametric probabilistic models and non-or semiparametric probabilistic models.

Fully parametric probabilistic models, including Poisson, negative binomial (NB) regression models and their variants, are extensively used in the studies of crash frequency over past decades.

2.2.1 Poisson Regression Models

The Poisson regression models are basic models used to analyze count data. Prior to using Poisson regression models, normal linear regression models with log transform of crash count data have been used in crash data analysis. However, Miaou and Lum (1993) pointed out underlying distributional assumption (i.e., normal distribution) of log-transformed regression models cannot adequately describe the discreteness and non-negativity of crash occurrence. In Poisson regression models, it is assumed that the response variable *Y* has a Poisson distribution, and the logarithm of its expected value can be modeled by a linear combination of unknown parameters. If y_i is the number of crashes per year for a particular site or roadway section *i* with corresponding vector $\mathbf{x_i}$ of the predictor variables, a basic Poisson regression model assumes the probability function of y_i to follow a Poisson distribution with the mean λ_i:

$$P(Y_i = y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y_i!} \quad (2.1)$$

The expected number of crashes λ_i is conditioned to the explanatory variables $\mathbf{x_i}$ through a log link function:

$$E(y_i) = \lambda_i = exp(\boldsymbol{x_i}'\boldsymbol{\beta}) \quad (2.2)$$

where $\boldsymbol{\beta}$ is the vector of parameters that can be estimated by maximum likelihood method.

The Poisson distribution has one parameter that simultaneously determines the conditional mean and variance. Therefore, the Poisson regression model as described above implies that the conditional mean function and variance function to be equal.

Miaou et al. (1992) explored the effects of roadway geometric design on truck crashes by a Poisson regression model with data from the Highway Safety Information System (HSIS). Kumara and Chin (2005) adopted a Poisson model to represent the reported crash counts and a probit model to describe the reporting mechanism of the crash happened at three-legged signalized intersections.

Other researchers also used Poisson regression models in their studies to capture discrete and non-negative traits of crash data (e.g., Abdel-Aty and Pemmanaboina, 2005; Kumara and Chin, 2005; Daniel and Chien, 2004; Ivan et al., 2000; Miaou and Lum, 1993).

The Poisson regression models are simple and robust. They preserve the original nature of crash counts (i.e., discreteness and non-negativity) without transforming them

into another scale. Despite of these advantages, a shortcoming of these models is that the variance is assumed to be equal to the mean – equi-dispersion, which is commonly not the case for crash count data.

2.2.2 Negative Binomial Regression Models

To relax the equi-dispersion constraints of the Poisson regression models, NB regression models are widely adopted. In the basic form of a NB regression model, if y_i is the number of crashes per year for a particular site or roadway section i with corresponding value $\mathbf{x_i}$ of the predictor variables, the negative binomial probability mass function of y_i is expressed as:

$$P(Y_i = y_i) = \frac{\Gamma(r + y_i)}{\Gamma(y_i + 1)\Gamma(r)} \left(\frac{r}{r + \lambda_i}\right)^r \left(\frac{\lambda_i}{r + \lambda_i}\right)^{y_i} \tag{2.3}$$

where Γ is the gamma function, and r is the dispersion parameter. The mean and variance of y_i are:

$$E(y_i) = \lambda_i = exp(\mathbf{x_i}'\boldsymbol{\beta}) \tag{2.4}$$

$$Var(y_i) = \lambda_i + \frac{{\lambda_i}^2}{r} \tag{2.5}$$

where $\boldsymbol{\beta}$ is the vector of parameters that can be estimated by maximum likelihood method. As the dispersion parameter r decreases, the variance of the negative binomial distribution increases. When the dispersion parameter r gets large (and λ_i is fixed), the

negative binomial converges to a Poisson distribution: this means that the negative binomial model is a more general model than the Poisson and it can be motivated as a mixture of Poisson distributions.

Shankar et al (1995) compared the Poisson model with the NB regression model in examining the effects of roadway geometric and environmental factors on the occurrence of freeway accidents. Miaou (2001) employed the NB regression model to estimate vehicle roadside encroachment rates for rural two-lane highways, considering AADT, lane width, horizontal curvature, and vertical grade as covariates. The Highway Safety Manual (HSM) 1st edition, which intends to "provide practitioners with the best factual information and tools to facilitate roadway design and operational decisions based on explicit consideration on their safety consequences" (ASSHTO, 2010), was published in summer 2010. In HSM, the base model for Rural Two-Lane Highways was developed with negative binomial regression analysis for data from 619 rural two-lane highway segments in Minnesota and 712 roadway segments in Washington State which was obtained from the FHWA HSIS (FHWA 2000). Other studies addressing application of NB regression models in crash analysis include Zegeer et al. (2001), Vogt and Bared (1998), etc.

One of the limitations of NB regression models is that their simple but strict variance assumptions, which are often violated by the crash data. In NB regression models, it is assumed that each individual has the same probability distribution (homogeneity assumption) and that they are independent (independence assumption). However, crash data often exhibits over-dispersion (Park and Lord, 2009), reflecting

heterogeneity or lack of independence among individuals. This is because that entities with the same represented traits have different means since the unrepresented traits (measured or unmeasured) are not included in the model (Hauer, 2001).

2.2.3 Zero-inflated Regression Models

Many empirical crash count data exhibit extra zero observations more that could be handled by the Poisson or NB regression models. Several studies have employed the zero-inflated regression models (i.e., zero-inflated Poisson (ZIP) models, zero-inflated negative binomial (ZINB) models) to describe this type of situation (Shankar et al., 1997; Shankar et al., 2003; Lee and Mannering, 2002; Chin, 2003). For example, Chin (2003) adopted a zero-inflated negative binomial (ZINB) regression model to address the problem of excess zero crash counts, which is an obvious manifestation of over-dispersion in crash data.

Zero-inflated regression models assume that the data is a mixture of two data generating processes: one generates zeros, and the other generates non-negative values through either a Poisson or a negative binomial data-generating process. The result of a Bernoulli trial can be used to determine which of the two processes generates an observation. In general the process can be described as shown below:

$$y_i \sim \begin{cases} 0 & with\ \ probability\ \varphi \\ g(y_i|\mathbf{x}_i) & with\ probability\ 1-\varphi \end{cases} \tag{2.6}$$

where y_i is the number of crashes per year for a particular site or roadway section i .

$\mathbf{x}_i$ is a vector of the covariates

φ is the probobility that process one is chosen.

$g(.)$ generates counts from either a Poisson or a negative binomial model.

Thus, the probability mass function of y_i is

$$P(Y_i = y_i|\mathbf{x}_i, \mathbf{z}_i) = \begin{cases} \varphi(\mathbf{z}_i'\boldsymbol{\theta}) + (1 - \varphi(\mathbf{z}_i'\boldsymbol{\theta}))g(0|\mathbf{x}_i) \\ (1 - \varphi(\mathbf{z}_i'\boldsymbol{\theta}))g(y_i|\mathbf{x}_i) \end{cases} \quad (2.7)$$

where $\mathbf{z}_i$ is the vector of zero-inflated covariates.

$\boldsymbol{\theta}$ is the vector of zero-inflated coefficients.

ZINB and other Zero-inflated counts regression models can fit the data well when the underlying assumption that the excess zeros solely explain the heterogeneity of data is valid; however, such an assumption might not always be true.

2.2.4 Other Models

In addition to the models reviewed in the previous sections, several other techniques in count data modeling have been applied to crash data analysis (Miaou and Lord, 2003; Hilbe, 2007; Mitra and Washington, 2007; Park and Lord, 2009). Hilbe (2007) introduced the varying dispersion parameter models, with which users can determine sources influencing over-dispersion. However, finding appropriate covariates that impact the over-dispersion can be problematic if the over-dispersion is partly due to unobserved variables. Park and Lord (2009) recommended an alternative formulation, finite mixture

regression models of Poisson or NB, to capture heterogeneity and handle over-dispersion, especially when the crash data is suspected to belong to different groups.

Lord et al. (2005) reviewed most commonly applied count data models in crash modeling, including Poisson, binomial, NB, Zero-Inflated Poisson and Negative Binomial regression Models (ZIP and ZINB), and Multinomial probability models, and also provided some guidance on how to select appropriate models for different conditions.

Despite all the advantages of fully parametric probabilistic count data models in terms of modeling, estimation and inference, there are certain drawbacks preventing reliance on these models as the predictive analysis tools. Fully parametric probabilistic models impose restrictive parametric assumptions regarding how covariates affect the response variable. As a consequence, those models usually lack of robustness, even when flexible models like mix models are applied. Moreover, those models could result in inconsistent or very poor estimation and inference under inappropriate distribution assumptions. The limitation of their ability to handle heterogeneity is another big concern. Given these limitations, it is attractive to study how non-or semiparametric models which freely approximate the conditional distribution perform for crash modeling.

2.3 Quantile Regression and Quantile Regression on Counts

Quantile regression (QR), as introduced in Koenker and Bassett (1978), extends the classical least squares estimation of conditional mean models to the estimation of a

range of models for conditional quantile functions. By complementing the exclusive focus of the conditional mean, QR offers a systematic strategy for examining how covariates influence the location, scale, and shape of the response distribution, and thus offers the opportunity for "a more complete view of the statistical landscape and the relationships among stochastic variables." (Koenker 2005) The mathematical forms of quantile regression are distinct from the method of least squares, where the squared errors are minimized. Instead, QR leads to problems in linear programming that can be solved by the simplex method. In recent years, research and applications of quantile regression can be found in various areas, including medical reference, survival analysis, financial economics, environmental modeling, ecology analysis and the detection of heteroscedasticity (Yu et al, 2003; Cade and Noon, 2003).

QR has been used in the context of linear regression models and other continuous regressions as an alternative to least squares for a long time, but its application to count data was not developed until very recently, because of the difficulty in obtaining valid asymptotic results for a non-differentiable objective function combined with a discrete response variable. Machado and Santos-Silva (2005) succeeded in extending QR to count data models through a "jittering" process that artificially imposes some degree of smoothness to overcome this problem. Implementation and applications of such a method can be found in Winkelmann, 2005; Miranda, 2008; Moreira and Barros, 2009. Their research shows the potential benefits of applying QR for counts to crash modeling.

QR for counts can provide crash modeling semiparametric technologies. It allows researchers to relax the distribution restriction on the response variable, allowing more

robust estimation (Miranda, 2008). Moreover, QR for counts does not restrict the way covariates affect different regions of the outcome distribution, and thus provides a more complete picture of the relation between explanatory variables and crash frequency. It is more appealing to investigate the "complete picture" rather than just the mean, considering conditional distributions of crash data usually skew to right. For example, it is possible to determine whether features such as speed limit and shoulder type have different effects for high risk segments towards those with low risk. Moreira and Barros (2009) pointed out "In the count world it is common that features other than location depend on the covariates, making the estimation of the conditional expectation poorer in the sense that provides very little information about the impact of the regressors on the outcome of interests", and it is "potentially interesting to study the effect of regressors not only on the mean but also on single outcomes and in the full distribution".

In this research, QR for counts is adopted as a methodological alternative to provide a fuller analysis on the relation between crash frequency and relevant factors, such as road geometry and traffic characteristics.

2.4 Summary

This chapter presented the literature review that forms the background of this research. The literature review started with a review on state of the art and practice in the area of developing road risk indices to evaluate the road safety performance in both U.S. and other countries around the world. The review shows that assessing safety performance of existing roads by composite indicators has demonstrated its practice

advance and is becoming an accepted practice at the international level. In the U.S., research in this area is very limited, and no applications have been found in practice. Since crash prediction models serve as a key for assessing road safety performance, a thorough review of different methods for traffic crash occurrence modeling as well as discussions on their advantages and disadvantages was presented in the second section.

Although numerous performance models are available for describing the relation between crash frequency and their explanatory factors, they suffer from various limitations. In the third part, quantile regression and its application to count data were reviewed. The review reveals that the technique of QR on counts has demonstrate its power for modeling count data in many fields and it has the potential to provide a fuller and more robust analysis of crash data compared to the traditional crash prediction models.

CHAPTER 3 METHODOLOGY

In order to achieve the objectives of the research, a comprehensive framework for developing RRIs of existing roads for both road users and agencies is proposed. The conceptual framework of the methodology along with the various modules is shown in Figure 3.1. First, the Highway Safety Information System (HSIS) dataset which provides both roadway characteristics and crash data is identified for this research. Second, a framework of formulating Road Risk Indices (RRIs) to capture the nature of road risks and reflect relative safety levels of the facilities in an objective and proactive manner is introduced. The formulation is able to incorporate the results from traffic crash prediction models the development of the RRIs. Third, a semi-parametric count model is discussed and quantile regression is employed for estimation to provide a fuller and more robust analysis on the relation between crash frequency and relevant factors. Compared to classical methods where only one set of point estimates and one prediction for each input vector are obtained, quantile regression based models give set of estimates and then a predicted conditional distribution for the response variable. Fourth, two general methods are provided to utilize the distributional information for point estimate of crash rates. The predicted crash rates are employed as the key part for developing RRIs. Fifth, a Geographic Information System (GIS) is proposed to provide a platform to store, manage, and present RRIs and related data with reference to geographic location data. The details of the framework are described in the following sections.

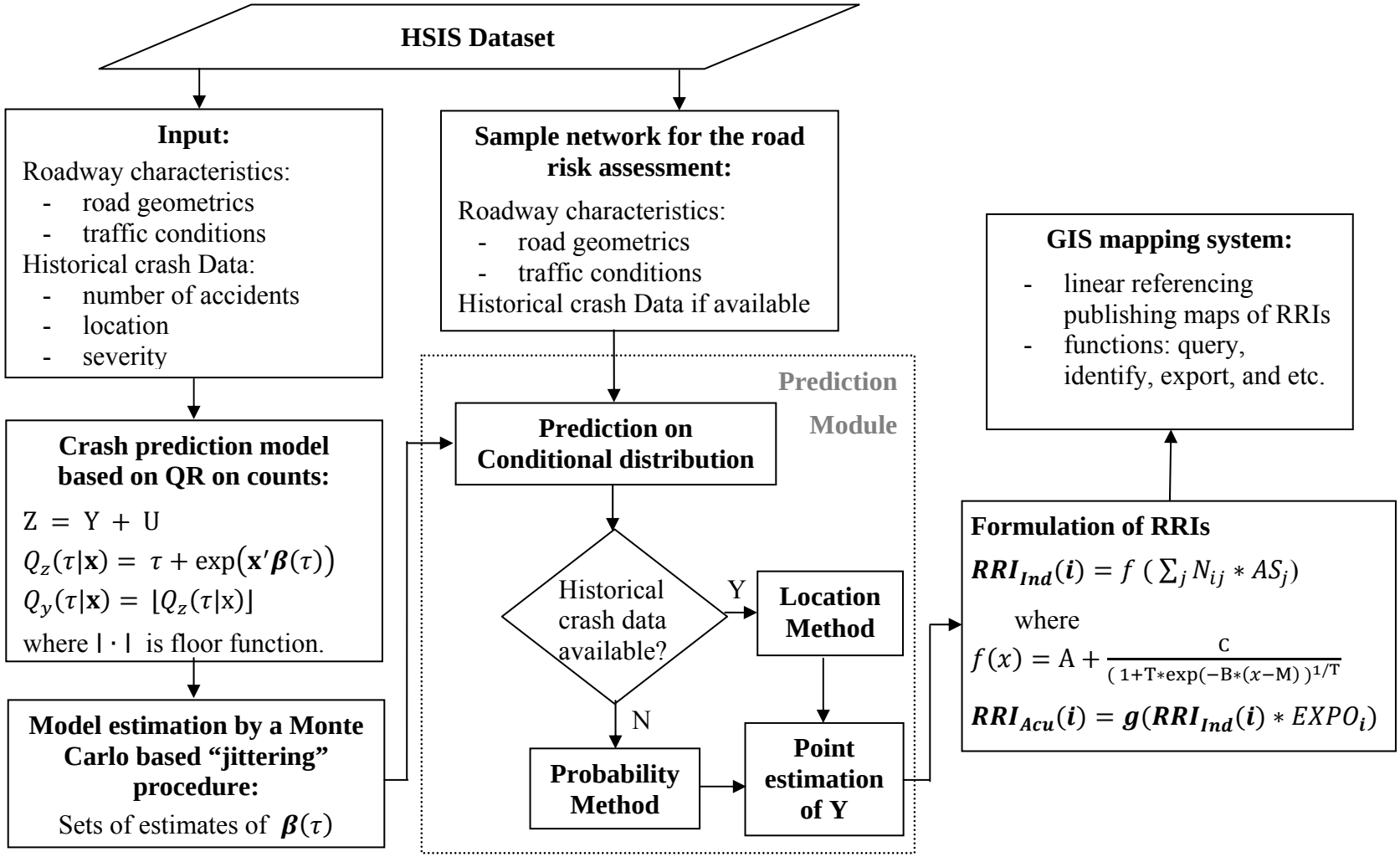


Figure 3.1: Major Components of the Methodological Framework

3.1 HSIS Program

The Highway Safety Information System (HSIS) operated by the University of North Carolina's Highway Safety Research Center (HSRC) and LENDIS Corporation under contract with the Federal Highway Administration (FHWA), is a multistate database containing crash, roadway inventory, and traffic volume data. The HSIS is used in support of the FHWA safety research program. Currently, there are nine selected States participating in the program, including California, Illinois, Maine, Michigan, North Carolina, Ohio, and Washington, as shown in Figure 3.2. "The participating States were selected based on the quality and quantity of data available, and their ability to merge data from various files."(HSIS, 2010)

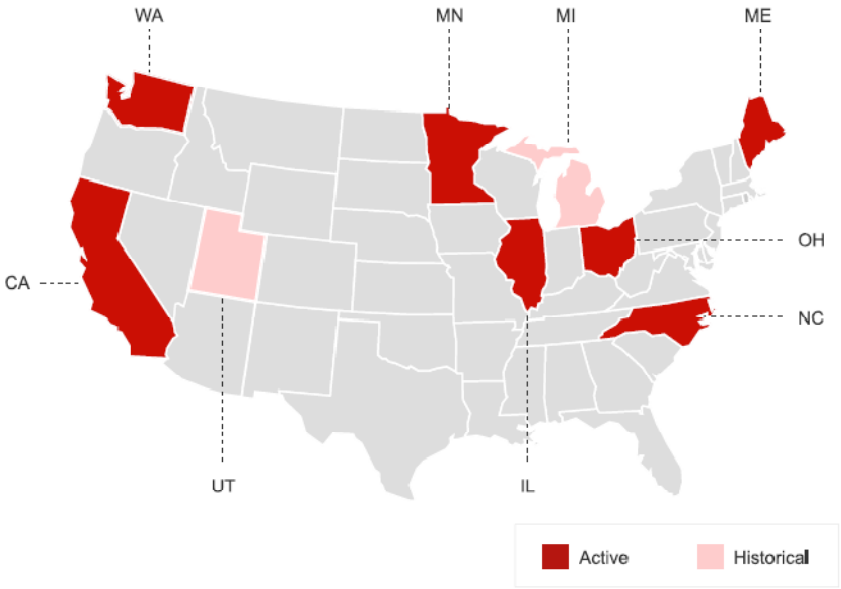


Figure 3.2: States Participating in the HSIS Program (HSIS, 2010)

While all of the selected States maintain basic crash files, roadway inventory files, and traffic files, individual States also collect other types of data. Among those States, all of them collect data for roadway segments, and two of them provide data for intersections. Within the HSIS, roads are divided into segments with homogenous characteristics (segments begin and end when certain geometric characteristics change). Detailed roadway segment information, such as horizontal curvature and grade attributes, surface width, lane width and type, shoulder width and type, median information and other variables, is available in the database. In addition, the HSIS contains only police-reported crash data on the State-maintained highway system, where all types of vehicle crashes occurring on each segment, including fatalities, injuries, property-damage-only, and other relevant information, are recorded in the database. As one of the most comprehensive databases of crash data, the HSIS database is selected for this research.

3.2 Formulation of RRIs

Although most of the current research and practices mainly focus on developing Road Risk Indices to serve as an aid tool for safety management agencies, ideally the interest of road users should also be considered when developing such indices. Wu and Zhang (2010) demonstrated that providing travelers with information on RRIs can significantly improve traffic safety with a small increase in travel time at the aggregated level. In this section, the formulation of RRIs for both road users and highway agencies is discussed.

Generally, road crashes and injuries are the results of a complex process, and the risk can be modeled by a three dimensional space of influencing factors, namely exposure, crash rate, and crash severity (Elvik and Vaa, 2004), as shown below:

$$\text{Road risk} = \text{function (exposure, crash rate, severity)} \tag{3.1}$$

Exposure denotes the amount of activity in which the crash may occur (Elvik and Vaa, 2004). The amount of activity is usually represented by the amount of travel which can be defined and measured in various ways. For instance, the number of vehicle miles traveled (VMT) and traffic volume are commonly used as the measurements for exposure.

Crash rate is the number of crashes per unit of exposure. It is proportional to the probability of crash occurrence and always regarded as the key for evaluating road risk.

Crash severity refers to the consequence of crashes in terms of injuries to people and damage to property. Although it should be a continuous variable, crash severity is always classified into several categories for practical use. For example, the official road crash statistics in the U.S. classify crashes by their severity into three categories: fatality, injury, and property damage crash.

In this research, two RRIs are proposed: 1) the index providing safety information for individual drivers (RRI_{Ind}) and 2) the index reflecting safety performance of roadway sections and intersections (RRI_{Acu}) for use by highway agencies. More specifically, the RRI_{Ind} is developed to measure an individual driver's risk of each exposure on a homogenous roadway segment or an intersection, whereas the RRI_{Acu} is used to

represent the accumulated risk potential of a link or a node in terms of its influence on the reliability of service a roadway segment or an intersection is supposed to provide..

3.2.1 Road Risk Index for Individual Exposure

Because RRI_{Ind} is intended to describe road risk per unit exposure, it is formulated as a function of two factors: crash rate and crash severity. The factor exposure is considered implicitly in the evaluation of crash rate.

$$RRI_{Ind}(i) = f\ (\sum_j N_{ij} * AS_j) \tag{3.2}$$

where i = roadway segment or intersection i;

j = 1 to 3, indicating three types of crashes: t fatality, injury, and property damage only;

N_{ij} = predicted number of type j crash on roadway segment i (per million vehicle-mile), or predicted number of type j crash at intersection i (per million vehicle);

AS_j = relative level of costs in corresponding to crash type j; and

$f\ (\cdot)$ = transformation function which maps the input to a desired range.

More specifically, to calculate N_{ij}, both the crash rates on roadway segment and the crash severity distributions are needed. While a quantile regression based crash model is developed for the prediction of crash rates, typical estimates are adopted as the crash severity distributions for roadway sections and intersections. For instance, HSM offers

the distribution of crashes by severity categories for different types of highways (HSM, 2010).

AS_j is derived from the estimated average crash cost by crash type. Based on a report from the Bureau of Transportation Statistics (BTS), the averaged costs of a fatality, injury, and property damage crashes are $4,113,956, $144,291, and $6,783, respectively. For this study, these values are normalized using the cost of a property damage crash as the basis, resulting in cost factors of 607, 21, and 1 respectively. These values are used as the AS_j in correspondence with each type of accidents.

RRI_{Ind} is defined on a scale of 0-10, with 0 representing the lowest road risk and 10 the highest. Various transformation functions can be applied to convert an input value to the desired range. In this research, a sigmoid function is proposed to perform this conversion. Stephens (1990) used the term risk to describe the situation when an individual has an awareness of the probability distribution of an event, and a utility function was usually used to describe the risk sensitivity of users of the information. Typical forms of utility functions include linear, concave, or possibly even convex functions. A sigmoid, or S-shaped, utility curve can model both risk taking and risk aversion because it has both convex and concave curves (Friedman & Savage, 1948). In the study conducted by Friedman and Savage, it is argued that an individual's sensitivity to risk would vary according to their wealth; at some levels an individual might take risks, and at other levels an individual might avoid them. Considering the modeling flexibility, sigmoid functions are used as the utility function in economics and human behaviors studies. Driver attitude towards crash risk is a complex phenomenon and

involves individual's differences (Golias et al., 1997; Kanellaidis et al., 1999). Yannis et al. (2005) developed a model to quantitatively describe driver reactions towards crash in Greece with the use of stated preference techniques. A logistic regression model was applied to identify the driver sensitivity parameters that influence his choices in order to reduce the crash risk. Results showed a sigmoid function performed well in that research. In this research, the sensitivity function of road users and transportation agencies towards road crashes is assumed to exhibit a sigmoid shape.

Commonly used sigmoid functions include standard logistic function, Gompertz function, Weibull function, Richards model, etc., where the logistic function is typically used for risk analysis. In this research, the generalized logistic function (Richards model) is adopted for two advantages it has in terms of its flexibility in modeling process: 1) allow a non-symmetrical model format, and 2) allow inflection point to vary.

Therefore, $f(\cdot)$ is defined as a generalized logistic function:

$$f(x) = \mathrm{A} + \frac{\mathrm{C}}{(1+\mathrm{T}*\exp(-\mathrm{B}*(x-\mathrm{M})))^{1/\mathrm{T}}} \tag{3.3}$$

where A = the lower asymptote (e.g.,0),

C = the upper asymptote (e.g.,10),

M = the time of maximum growth,

B = the growth rate, and

T = near which asymptote maximum growth occurs.

The assumption and the form of the transformation function are made based on the state-of-the-art on this topic. Other forms of sensitivity functions that can describe the risk sensitivity function in a better way might exist, and deserve further research.

3.2.2 Road Risk Index for Roadway Sections

For a homogenous road segment, all of the three factors are considered in developing the RRI_{Acu} which is defined as

$$RRI_{Acu}(i) = g\,(RRI_{Ind}(i) * EXPO_i) \tag{3.4}$$

where $EXPO_i$ = exposure in million vehicle-miles of travel per year on roadway section *i*: $ADT_i * L_i * 10^{-6} * 365$, or exposure in million vehicle-miles of travel per year in intersection *i*:$\left(ADT_{i,1} + ADT_{i,2} \right) * 10^{-6} * 365$

ADT_i = average daily traffic (ADT) for homogenous segment *i*,

$ADT_{i,1}$ = average daily traffic on the major road at intersection *i*,

$ADT_{i,2}$ = average daily traffic on the minor road at intersection *i*, and

L_i = length of homogenous segment *i*.

$g\,(\cdot)$ = transformation function which maps the input to a desired range.

3.3 Quantile Regression

Quantile regression (QR), originally developed by Koenker and Bassett (1978), is an extension of the classical least squares estimation of conditional mean models to the

estimation of models for conditional quantile functions. To briefly recall the ordinary quantile, consider a random variable Y with probability distribution function

$$F(y) = P(Y < y) \tag{3.5}$$

Then for any $\tau \in (0,1)$, the τth quantile of Y is defined as

$$Q(\tau) = \inf\{y : F(y) \geq \tau\} \tag{3.6}$$

Quantile regression estimates the value of the parameter vector $\boldsymbol{\beta}(\tau)$ from the set of allowable vectors that minimize the loss function

$$\hat{\beta}(\tau) = argmin_{\beta \in \boldsymbol{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \xi(\boldsymbol{x}_i, \boldsymbol{\beta})) \tag{3.7}$$

where $\rho_\tau(z) = z\big(\tau - I(z < 0)\big), 0 < \tau < 1$, and $I(\cdot)$ is the indicator function which is one when the argument is true and zero otherwise. The general τ^{th} sample quantile $\xi(\boldsymbol{x}, \boldsymbol{\beta}(\tau))$ is the analogue of $Q(\tau)$. When $\xi(\boldsymbol{x}, \boldsymbol{\beta}(\tau))$ is formulated as a linear function of parameters, the above minimization problem can be solved very efficiently by linear programming methods (Koenker, 2004). If the model fits well, a plot of fitted versus actual values will show that τ percentage of observed values are less than the fitted values, while 1 − τ percentage of the observed values are greater than the fitted values.

3.4 Quantile Regression for Counts

To analyze crash data using quantile regression models, the methodology of employing QR on counts through a "jittering" process developed by Machado and Santos-Silva (2005) was adopted in this research. In this section, a general review of such a methodology and its implementation is presented.

As briefly discussed earlier, the main problem for QR on counts is its non-differentiable objective function combined with a discrete response variable. The problem regarding non-differential objective functions, which is always the case for any QR model, has been well solved by introducing linear programming and the simplex method. In order to overcome the other difficulties, Machado and Santos-Silva (2005) introduced a "jittering" procedure to smooth the dependent variables.

Let Y be a count random variable, $\mathbf{X}$ be the vector of the covariates. Then the τ^{th} quantile of Y can be defined as

$$Q_Y(\tau) = \min\ (\eta \mid P(Y \leq \eta) \geq \tau) \qquad \text{where} \qquad 0 < \tau < 1 \qquad (3.8)$$

Since Y has a discrete distribution, $Q_Y(\tau|\mathbf{x})$, which denote the τ^{th} quantile of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$, cannot be a continuous function of the parameters of interest. To overcome this limitation, Machado and Santos Silva (2005) constructed a continuous random variable Z as the sum of Y and a uniform random variable U, which is independent of Y and $\mathbf{X}$.

$$Z = Y + U \qquad \text{where} \qquad U \sim uniform[0,1) \qquad (3.9)$$

For general count data, let $P(Y = k|\mathbf{x}) = p_k$, and the conditional probability density function of Z can be written as

$$f(z|\mathbf{x}) = p_k \qquad \text{for} \qquad k \leq z < k+1 \tag{3.10}$$

The conditional cumulative distribution function of Z can be written as

$$F(z|\mathbf{x}) = \begin{cases} p_0 z & for \quad 0 \leq z < 1 \\ \sum_{i=0}^{k-1} p_i + p_k(z-1) & for \quad k \leq z < k+1 \end{cases} \tag{3.11}$$

Hence, the conditional quantile function of Z can be written as

$$Q_z(\tau|\mathbf{x}) = \begin{cases} \dfrac{\tau}{p_0} & for \qquad 0 \leq \tau < p_0 \\ k + \dfrac{\tau - \sum_{i=0}^{k-1} p_i}{p_k} & for \quad \sum_{i=0}^{k-1} p_i \leq \tau < \sum_{i=0}^{k} p_i \end{cases} \tag{3.12}$$

This transformation ensures that the quantiles of Z are non-negative and continuous in τ. In addition, such a transformation is linear in the parameters of the covariates (Moreira and Prato, 2009). Machado and Santos-Silva (2005) also indicated that "using the uniform noise to jitter the data is by no means a necessity", and other smoothing noise may be generated by any "continuous distribution with support on [0, 1) and a density bounded away from 0".

After constructing the continuous random variable Z, Machado and Santos-Silva (2005) recommended that the τ^{th} quantile of Z be modeled by the following formulation:

$$Q_z(\tau|\mathbf{x}) = \quad \tau + \exp\left(\mathbf{x}'\boldsymbol{\beta}(\tau)\right) \tag{3.13}$$

The reason of adding τ on the right side of equation (3.13) is to impose a lower bound of τ to $Q_z(\tau|\mathbf{x})$, which is bounded from below by τ due to the way it is constructed. The exponential form is included to keep in line with traditional forms in count data models. Next, transform Z as follows:

$$Q_{T(z;\tau)}(\tau|\mathbf{x}) = \quad \mathbf{x}'\boldsymbol{\beta}(\tau)$$

$$\text{where } T(z;\tau) = \begin{cases} \log(z-\tau) & \text{for} \quad z > \tau \\ \log(\delta) & \text{for} \quad z \leq \tau \end{cases} \tag{3.14}$$

with δ a suitably small positive number ($0 < \delta < \tau$). Then, the parameters of covariates $\boldsymbol{\beta}(\tau)$ can be estimated by minimizing an asymmetrically weighted sum of absolute errors

$$\min_{\beta \in \mathrm{R}^k} \sum_{i=1}^{n} \rho_\tau\left(T\left(z_j;\tau\right) - \mathbf{x}_i'\boldsymbol{\beta}\right) \tag{3.15}$$

$$\text{where} \qquad \rho_\tau(v) = v(\tau - I(v < 0))$$

Although the quantile function is not differentiable everywhere, the set of those corner points will have measure zero if there is at least one continuous covariate and the general valid asymptotic inference still holds. More specifically, Machado and Santos-Silva (2005) proved that the estimator is consistent and asymptotically normal distributed. Thus, typical tests, like the Wald test and t-test, can be used to inference the

results. Also, since a uniform "noise" has been artificially added, a Monte Carlo based "average-jittering" procedure was suggested to average it out and make the estimator more efficient and robust.

The final question is how to interpret the parameters. For users, the interest lies in how the covariates affect $Q_y(\tau|\mathbf{x})$, not $Q_z(\tau|\mathbf{x})$ in many cases. According to equation (3.9), $Q_y(\tau|\mathbf{x})$ can be recovered from $Q_z(\tau|\mathbf{x})$ by using $Q_y(\tau|\mathbf{x}) = \lfloor Q_z(\tau|\mathbf{x}) \rfloor$, where$\lfloor \cdot \rfloor$ donates floor function. The impact of a change of explanatory variable x_j from $x_j{}^0$ to $x_j{}^1$ on the conditional quantile of *y*, given all the other covariates remain the same, can be calculated by

$$\Delta Q_y = \lfloor Q_z(\tau|\mathrm{x}_j{}^0, \mathbf{x}) \rfloor - \lfloor Q_z(\tau|\mathrm{x}_j{}^1, \mathbf{x}) \rfloor \tag{3.16}$$

Since one can obtain $Q_y(\tau|\mathbf{x})$ from $Q_z(\tau|\mathbf{x})$, but not vice versa, special caution should be taken when one interprets the meaning of the coefficients. If a null hypothesis $\beta_j(\tau) = 0$ is not rejected, it means that the coefficient of x_j at τ^{th} quantile is not significantly different from zero and that the impacts of x_j on $Q_z(\tau|\mathbf{x})$ and consequently $Q_y(\tau|\mathbf{x})$ are not significant. On the other hand, if the null hypothesis is rejected, it means the impact of x_j on $Q_z(\tau|\mathbf{x})$ is significant, but the changes in x_j may or may not affect $Q_z(\tau|\mathbf{x})$. In fact, a change of x_j will only affect $Q_y(\tau|\mathbf{x})$ only if the change is able to change the integer part of $Q_z(\tau|\mathbf{x})$. Machado and Santos-Silva called it "magnifying glass effect" on $Q_z(\tau|\mathbf{x})$, and pointed out that it is "easier to detect dependence of the distribution of Y on **X** by looking at $Q_z(\tau|\mathbf{x})$ than by looking at $Q_y(\tau|\mathbf{x})$."

3.5 Crash Prediction Using Conditional Distribution

As discussed earlier, there are advantages of QR on counts over the traditional models especially with the presence of heterogeneity, and the method can provide a fuller picture and more robust estimate of crash numbers. However, since the results from quantile regression are a set of estimates of parameters rather than one, how to take advantage of such distribution information to yield better point prediction raises a challenging issue. Actually, the more the information, the more challenging it is to combine everything in a single index.

On the other hand, due to data availability and limitations of knowledge and understandings about the process of traffic accidents, crash prediction models certainly cannot take all the influencing factors of road safety into consideration. Therefore, roadway features are not the only clue to estimating the safety of an entity, because historical crash records can also help. Thus, it would be wise to build a theoretical framework that can combine the information contained in historical crash records with that obtained from the crash prediction models, if the historical crash data is available.

Although limited literature has been found addressing the above issues, Ma and Pohlman (2008) gave thorough discussions on the topic. The paper applied quantile regression to return forecasting and portfolio construction of securities in the context of financial markets. Two general methods, namely quantile regression alpha distribution (QRAD) and quantile regression portfolio distribution (QRPD), where the former used conditional distributional information at the forecasting stage and the latter at the portfolio construction stage, were proposed. Proof of properties of these methods was

also presented in the paper. The results showed that their proposed methods provide more accurate forecasts and potentially higher value-added portfolios. In this research, methodology of QRAD is adopted and revised to accommodate the needs of constructing RRIs.

From the QR on counts model, predictions on the distribution of the number of crashes F(Y) given **X** (e.g., roadway geometrics, road side features, and traffic conditions) can be obtained. To utilize the estimated conditional distribution to predict the number of accidents, two methods namely the Location Method and the Probability Method, are proposed in this section. While the location method deals with the situation where historical crash data of an entity is available, the probability method is used to predict the number of crashes entirely based on roadway features **X**.

3.5.1 Location Method

The Location Method assumes that the safety conditions of the entities remain in the same quantiles from one period to another. The method is based upon the assumption that for the most part of the time, the rank of safety does not change dramatically. By the Location Method, the prediction for individual roadway entity is determined in two steps: 1) determine the quantile of this entity by its historical crash data and the predicted conditional distribution, and 2) estimate the number of crashes for the selected quantile level. Thus for sets of road entities that are in the same quantile, they have same predictions.

More specifically, let τ = (0.25, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95), and the corresponding $\hat{Q}_y(0.25|\mathbf{x})$, $\hat{Q}_y(0.5|\mathbf{x})$, $Q_y(0.6|\mathbf{x})$, $\hat{Q}_y(0.7|\mathbf{x})$, $\hat{Q}_y(0.8|\mathbf{x})$, $\hat{Q}_y(0.9|\mathbf{x})$, and $\hat{Q}_y(0.95|\mathbf{x})$ can be obtained from the QR model. In this case, the predicted conditional distribution is divided into four intervals and the prediction of the number of crashes given **x** can be obtained as follows:

$$\hat{Y} = \begin{cases} \hat{Q}_y(0.25|\mathbf{x}), & \bar{Y}_h \leq \hat{Q}_y(0.5|\mathbf{x}) \\ \hat{Q}_y(0.6|\mathbf{x}), & \hat{Q}_y(0.5|\mathbf{x}) \leq \bar{Y}_h \leq \hat{Q}_y(0.7|\mathbf{x}) \\ \hat{Q}_y(0.8|\mathbf{x}), & \hat{Q}_y(0.7|\mathbf{x}) \leq \bar{Y}_h \leq \hat{Q}_y(0.9|\mathbf{x}) \\ \hat{Q}_y(0.95|\mathbf{x}), & \bar{Y}_h \geq \hat{Q}_y(0.9|\mathbf{x}) \end{cases} \tag{3.17}$$

where $\bar{Y}_h$ is the average of historical crash counts.

3.5.2 Probability Method

When there is no historical crash data available, the Probability Model should be used. Using the Probability Model, the estimated number of crashes given **x** can be calculated as follows:

$$\hat{Y} = \sum_{i=0}^{k} p_i \, \hat{Q}_y(\tau_i|\mathbf{x}) \tag{3.18}$$

where k is the number of divided intervals, and p_k is the probability of the occurrence of $\hat{Q}_y(\tau_k|\mathbf{x})$. For instance, using the same division as that of the Location Method, the predicted $\hat{Y}$ can be calculated as follows:

$$\hat{Y} = 0.5\hat{Q}_y(0.25|\mathbf{x}) + 0.2\hat{Q}_y(0.6|\mathbf{x}) + 0.2\ \hat{Q}_y(0.8|\mathbf{x}) + \ 0.1\hat{Q}_y(0.95|\mathbf{x}).$$

Ma and Pohlman (2008) proved that under mild conditions, both the Location Method and the Probability Method provide better goodness of fit than the conditional mean prediction. The relative accuracy of the predictions can be described by the following proposition: Let $\hat{Y}_i^{l,p}(\tau|\mathbf{x})$ be the composite quantile prediction for entity i from either the Location Method or the Probability Method, $\hat{Y}_i^{e}(\cdot\,|\mathbf{x})$ be the mean prediction, respectively, then

$$\sum_{i=1}^{N} |\hat{Y}_i^{l,p} - Y_i| \leq \sum_{i=1}^{N} |\hat{Y}_i^{e} - Y_i| \ . \tag{3.19}$$

Although the above proposition stands for situations with different number of divisions, there is a trade-off between the number of divisions, the accuracy of forecast, and the likelihood that an entity remains in the same quantile from period to period.

Since crash rate, rather than crash counts, is used in formulating RRIs, a further step is needed to convert the estimated number of crashes for a particular road entity to the predicted crash rate. The conversion is straightforward and can be described as follows:

$$N_i = \frac{Y_i}{L_i * \ AADT_i * 365} * 10^6 \tag{3.20}$$

where N_i is crash rate of entity i;

Y_i is number of crash counts for entity i;

$AADT_i$ is the average annual daily traffic for road segment *i*.

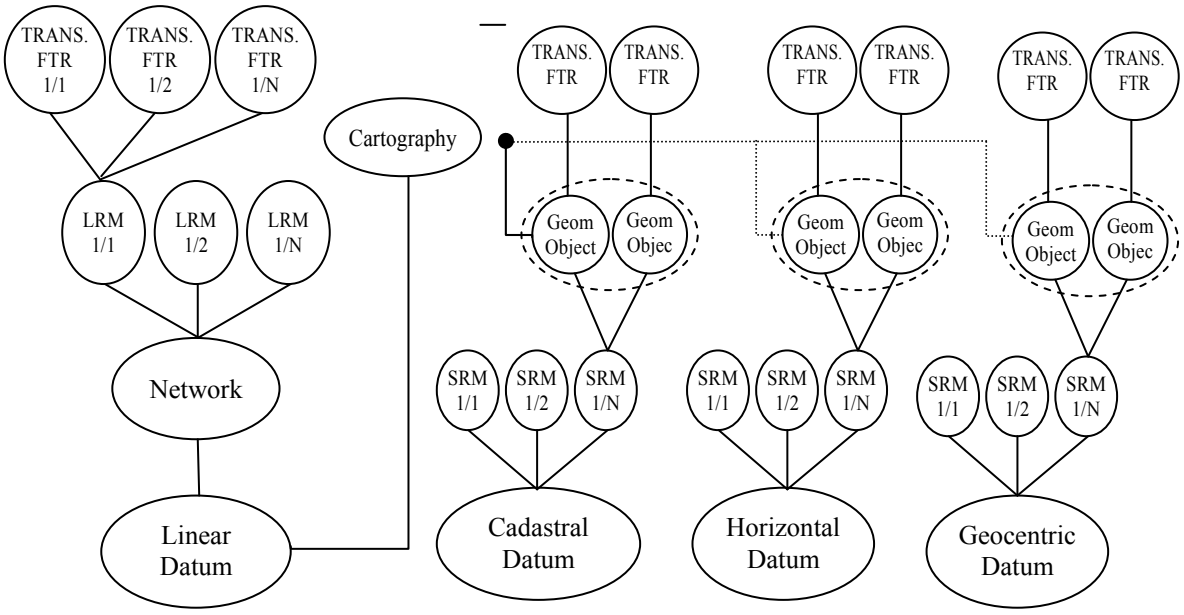
3.6 GIS Publishing System

GIS is a popular and powerful tool to visualize data, provide the capability to analyze data in relation to its location, and thus increase the value and utility of the information. In order to provide road users and highway agencies a user friendly interface to access the information, a Geographic Information System (GIS) platform is proposed to store, manage, and present RRIs and related data with reference to geographic location data. ESRI GIS packages including ArcGIS and its Linear Referencing tools can be used to develop the system. This publishing system should be able to provide basic GIS functions, including viewing and querying RRIs on the highway network, performing spatial analysis tasks such as selecting and buffering features, and manipulating information.

A Linear Referencing System (LRS) is a reference system in which the locations of features (points or segments) are identified by a relative measure along a linear feature. A location referencing method (LRM) is a way to identify a specific location with respect to a known point (Baker and Blessing, 1974). In a Linear Referencing System, each feature is located by either a point event (i.e., an intersection, or a crash site) or a linear event (a road segment). LRS is widely used for two primary reasons: 1) many locations are recorded as events along linear features; and 2) it can be used to associate multiple sets of attributes to portions of linear features without requiring that underlying lines are split each time attribute values change (ESRI, 2009).

Multidimensional Linear Referencing System (MDLRS), which was proposed in the National Cooperative Highway Research Program (NCHRP) 20-27 study, is a

commonly used conceptual framework in developing a Linear Referencing System in current practice. Figure 3.3 shows the conceptual model of the data in the MDLRS data model. The idea is to separate location referencing methods (LRMs) and geometric representations of the road network into separate tables, and link them through a temporally linear datum. This approach would provide a cohesive LRS that has capabilities to combine multiple location referencing methods, multiple cartographic representations, and multiple topological representations in an integrated system (Adams et al, 1997).



Note
LRM = Location Referencing Method
N = Total Number of LRMs per SRM
SRM = Spatial Referencing Method
TRANS. FTR = Transportation Feature

Figure 3.3: Conceptual Framework of the MDLRS Data Model (Adams et al, 1997)

For highway safety analysis, Linear Referencing is a viable method to aggregate data from different sources into an integrated dataset. For example, while traffic crash

and related information are usually recorded as a point event along a highway, roadway inventory are commonly stored as a segment event. In the HSIS dataset, crash data and roadway related data have been already aggregated using Linear Referencing techniques. The integrated data serves as the basic database for estimating crash prediction model and RRIs. Linear Referencing is the key technique for presenting the estimation results on maps as well.

3.7 Summary

This chapter discussed the fundamental methodology for developing RRIs to assess the safety performance of existing roads for both road users and agencies. The details of the main components for formulating and publishing RRIs, including the dataset for this research, the formulation of RRIs, the crash prediction model and its estimation method, techniques and tools to present the estimated RRIs with reference to geographic location data , were presented and explained. The next two chapters give more specific discussions on how to apply the proposed methodology to historical crash datasets.

CHAPTER 4 NUMERICAL ANALYSIS OF CRASH PREDICTION MODEL

In this chapter, a comprehensive case study was undertaken to demonstrate the application of the proposed methodology and models. The Highway Safety Information System (HSIS) dataset which provides both roadway characteristics and crash data was identified for this research. More specifically, yearly collected data from the Interstate Highway system in Washington State in the year 2002 were used to develop the crash prediction models. The highways were first divided into two groups considering the quality of crash data collected on them: one group of Interstate Highways in urban areas and the other group of those in rural areas. For each group, the abstracted dataset were further divided into two parts: one for developing the crash prediction model and the other for model validation and RRIs evaluation.

4.1 Dataset for Empirical Analysis

Data collected on the Interstate Highway system in Washington State in the year 2002 were used for the empirical study in this dissertation.

4.1.1 Washington State HSIS Database and Data Quality

The Washington State data in the HSIS database is derived from the Washington Transportation Information and Planning Support System (TRIPS), which is maintained

by the Transportation Data Office (TDO) of the Washington DOT. The Washington TDO provides the data to the HSIS program in the form of nine different data files, including:

- crash data,
- vehicle related data,
- occupant data,
- basic roadway inventory data,
- curve/grade/features data,
- roadway crossings and roadside facilities ("left/right") data,
- special-use lane information,
- railroad grade crossing index, and
- traffic data.

Crash related files include the Crash Subfile, the Vehicle Subfile, and the Occupant Subfile. The Crash Subfile contains over 75 variables which cover all basic variables that would be expected on standard police forms.

The basic Roadlog file contains information on such road features as surface width, lane width and type, shoulder width and type, median information, rural/urban codes, terrain codes, and various roadway type descriptors. The curve/grade/features file includes supplemental files on horizontal and vertical alignments, as well as roadway features such as bridges, tunnels, and underpasses. The Left/Right file contains information on crossing roadways, milepost markers, and roadside features such as rest areas and weigh stations. The Special-Use Lanes file contains information on bike lanes,

climbing lanes, etc. The Railroad Grade Crossing Index file contains information for each milepost where a railroad crosses a state route.

The Traffic Data file contains information for traffic count data, including AADT, single-trailer truck percentage, and double-trailer truck percentage. While features in the Roadlog, Curve, Grade and Ramp Files are section based, features in other files are point based, describing characteristics at a given milepost. Most of those files have been converted to section based files for ease of use in the HSIS system and crash analysis (Council and Mohamedshah, 2009).

The HSIS database for Washington State contains crash data for 1993 to 1996 and 1999 onwards. The 1997 and 1998 crash data were not included by WSDOT in the TRIPS system due to budget constraint problems encountered. Crash data was collected statewide by all police departments with a standard form. The preset crash reporting threshold is $750 for property-damage-only or personal injury. This criterion is generally used for crashes that occur on state highways in the rural areas. The completed standard forms are sent to Washington State Patrol (WSP) Crash Reports Division and the Washington TDO for location coding and other coding procedures. According to the TDO's analysis, the location coding for these state-route crashes is probably as accurate as it could be found in any state in the U.S., "with over 95% of the rural crashes being located to at least the nearest 1/10 of a mile, and over 95% of the urban crashes being located within 100 feet" (Council and Mohamedshah, 2009). While the physical reference markers for rural state systems are generally intact, some reference markers may be missing in some urban areas. Therefore, the accuracy of the location data for

crashes happened in rural areas is higher than that for crashes happened in urban areas on average.

In the HSIS database for Washington State, roadway-related data containing current characteristics of the state road system is derived from a series of roadway inventory files from their TRIPS system. In the database, there is information on approximately 7,000 center-line miles of mainline roadway and approximately 1,000 additional miles of ramp, frontage road, and other non-mainline roadway. All functional classes of roads within the state system (i.e., Interstate, U.S. and state routes) are included. Currently, road inventory data is available for 1993 to 1996 and from 2002 onwards. The 1997 to 2001 road inventory data were not included by WSDOT in the TRIPS system due to budget restriction problems they encountered.

4.1.2 Interstate (IS) Highway System in Washington

In Washington State, the Interstate Highway system consists of three primary routes (I-5, I-90, and I-82) and four auxiliary routes (I-182, I-205, I-405, and I-705), with total length of 764.27 center-line miles (WSDOT, 2009). I-90 is the longest primary Interstate Highway in Washington State measuring 297.52 miles, while I-5 is the second longest at 276.62 miles and I-82 is the shortest at 132.57 mi. The longest auxiliary Interstate Highway in Washington State is I-405 at 30.30 miles, and the shortest is I-705 at 1.50 miles. The other two auxiliary routes are I-182 at 15.19 miles, and I-205 at 10.57 miles.

In this research, yearly collected data for the Washington Interstate Highway system was extracted from the HSIS database and used for model development and validation. Considering the quality of crash data collection process, the highways were first divided into two groups: one group of Interstate Highways in urban areas and the other group of those in rural areas. For each group, the abstracted dataset were further divided into two parts: one for developing crash prediction model, including data collected on I-5, I-90, I-182, I-205, I-405, and I-705; the other for model validation including data collected on I-82. Data collected on I-82 was also used for model validation and RRIs evaluation discussed in both this and next chapters. Variables serving for the research interests and objectives were included; others were dropped from the original HSIS dataset. Also, segments with missing or inconsistent information were filtered out.

4.2 Estimation of Crash Prediction Model for Urban IS Highways

4.2.1 Data Description

The dataset used in developing the crash prediction model for Interstate Highways in urban areas contains 3934 highway segments with an average length of 0.064 miles. The total number of reported crashes of all types is 8,539 and for each individual segment the crash rate ranges from 0 to 49. The full descriptive statistics of the response variable and the covariates are presented in Table 4.1.

Table 4.1: Summary Statistics of Variables for Interstate Highways in Urban Areas in Washington State (3934 road segments)

Variable	Mean	Std. Dev.	Min	Max
Dependent variable				
Number of motor vehicle crashes (in 2002)	2.171	3.887	0	49
Covariates				
Geometric characteristics				
Road segment length (miles)	0.064	0.073	0.000114	0.76
Horizontal curve length (feet)	474.366	820.768	0	4631
Degree of horizontal curvature (°/100ft)	0.625	1.510	0	38.2
Vertical curve length (feet)	700.958	734.503	0	6000
Vertical grade (%)	1.518	1.315	0	6.03
Median barrier width (feet)	47.069	46.123	1	450
Number of lanes	5.614	1.429	3	9
Average lane width (feet)	12.721	2.515	11	39
Average shoulder width (feet)	8.866	3.335	0	17
Indicator for shoulder type curb (1 if yes, 0 otherwise)	0.039	0.193	0	1
Indicator for shoulder type wall (1 if yes, 0 otherwise)	0.079	0.270	0	1
Traffic characteristics				
Average annual daily traffic (AADT) per lane	18588	8130	2926	55677
Speed limit (mile/h)	61.843	4.324	30	70

In the dataset, since all the segments are homogenous in terms of main geometric characteristics, each segment should fall in the category of either straight sections or horizontal/vertical curves. On the other hand, a straight road or a curve can be divided into several segments based on other changes in road geometry. Thus, the covariate road segment length in the model stands for the length of a homogenous segment, a stretch of either a straight roadway or a curve. The covariates “horizontal curve length” and

"vertical curve length" represent the length of the whole horizontal/vertical curve in which an individual segment is located. Two dummy variables are included as indicators of the shoulder type. The default shoulder type, where indicators for shoulder type as curb or wall are both 0, is leveled shoulder.

4.2.2 Model Estimation and Results Analysis

The estimation was implemented using the *qcount* package of STATA (Miranda, 2006), which estimates quantile regression models for count data using the "jittering" method suggested by Machando and Santos Silva (2005). To select which quantiles to analyze, both the crash data at hand and practical interests were taken into consideration. The marginal distribution of crash data shows that about 40 percent of the segments had zero crashes in the year 2002. Therefore, it would be more convincing to analyze the conditional quantiles on the upper tail of the distribution, noting that the variation in the lower tail might be just due to random noise added in the jittering procedures. In practice, it is more interesting to look at segments with relatively large numbers of crashes, which are usually identified as black spots in road safety management. Thus, the remaining discussion mainly focuses on larger quantiles, even though the estimates of the first quantile in the results for the purposes of integral presentation are included. In addition, regarding the question of how to determine the number of jittering procedures, preliminary experiments have been carried out for each selected quantile. The results show that with 900 repetitions or more, there is no significant change of the estimates for

the selected quantiles when the number of repetitions is increased, and thus 900 repetitions was selected for the jittering procedures.

A negative binomial regression model was also selected as the benchmark of comparison in this research. A preliminary study was carried out to compare the Poisson and negative binomial models, and the results confirm the superiority of the latter. The estimate dispersion parameter is about 0.74 and the likelihood-ratio chi-square test returns 2914.28 with an associated *p*-value of 0.000, suggesting that the independent variable is over-dispersed and is not sufficiently described by the Poisson regression.

Estimation results are summarized in both Table 4.2 and Figure 4.1. Table 4.2 shows the parameter estimates and the related *z*-statistics (in the brackets) of the QRs on counts. Results for the first quantile, the median, each decile after the median, and the 0.95 percentile are included in the table. The coefficients and z-statistics in the last column are estimates from the NB regression model. Instead of *t*-statistics, STATA reports *z*-statistics and *p*-value to test the null hypothesis that the coefficient is equal to zero for count models. The test statistic *z* is the ratio of the coefficient to the standard error of the covariate, which follows a standard normal distribution.

Figure 4.1 presents a summary of quantile regression and NB regression results for urban Interstate Highway segments. The dashed line in each figure shows the NB regression estimate of the coefficients. The two dotted lines represent conventional 95 percent confidence intervals for NB regression estimate. The solid black line represents the quantile regression estimates of the coefficients. The shaded gray area depicts a 95 percent confidence band for the quantile regression estimates.

As shown in Table 4.2 and Figure 4.1, most of the covariates which are significant in the negative binomial regression model also tend to be significant in the proposed models. Among the geometric characteristics covariates, segment length, length of the horizontal curve, grade of vertical curve, median barrier width, number of lanes, shoulder width, and shoulder type are significant for most of the QRs and the NB model. For regressors that control traffic characteristics, the average annual daily traffic (AADT) per lane is significant under all the models, while the speed limit does not show any significant effect on crash frequency for this dataset at 1%. Moreover, the signs of the effects of the covariates that are significant do not switch at different quantiles in this study, which could not be the case for other datasets.

Table 4.2: Parameter Estimates for QRs and NB Model for Urban Interstate Highway Segments (z-statistics in brackets)

Variable	QR							Negative Binomial
	$Q_z(0.25\|x)$*	$Q_z(0.5\|x)$	$Q_z(0.6\|x)$	$Q_z(0.7\|x)$	$Q_z(0.8\|x)$	$Q_z(0.9\|x)$	$Q_z(0.95\|x)$	
Road segment length (miles)	10.349 [15.79]a	10.981 [24.91]a	11.071 [22.2]a	11.028 [20.72]a	10.279 [18.56]a	9.943 [16.49]a	9.217 [15.44]a	9.257 [29.78]a
Horizontal curve length (feet)	-8.88E-05 [-1.79]	-1.02E-04 [-2.43]b	-1.31E-04 [-2.86]a	-1.55E-04 [-4.77]a	-1.77E-04 [-5.03]a	-1.60E-04 [-4.32]a	-1.55E-04 [-4.32]a	-1.45E-04 [-5.16]a
Degree of curvature (°/100ft)	-0.006 [-0.23]	0.015 [0.38]	0.033 [0.71]	0.048 [3.18]a	0.034 [2.14]b	0.028 [1.55]	0.029 [1.48]	0.014 [0.90]
Vertical curve length (feet)	-9.95E-05 [-1.67]	-7.43E-05 [-1.51]	-5.88E-05 [-1.41]	-4.89E-05 [-1.19]	-5.07E-05 [-1.38]	-5.98E-05 [-1.75]	-8.97E-05 [-2.07]b	-5.73E-05 [-1.74]
Vertical grade (%)	0.076 [2.50]b	0.095 [4.35]a	0.092 [3.87]a	0.095 [4.70]a	0.082 [4.14]a	0.085 [3.87]a	0.133 [5.70]a	0.09 [5.65]a
Median barrier width (feet)	-1.63E-03 [-2.00]b	-2.13E-03 [-3.27]a	-2.38E-03 [-3.71]a	-2.47E-03 [-3.39]a	-2.20E-03 [-4.96]a	-2.07E-03 [-4.14]a	-1.83E-03 [-3.37]a	-2.24E-03 [-5.16]a
Number of lanes	0.266 [9.78]a	0.245 [12.54]a	0.251 [11.88]a	0.241 [10.11]a	0.225 [10.41]a	0.216 [11.33]a	0.196 [9.85]a	0.226 [15.45]a
Average lane width (feet)	0.003 [0.10]	0.008 [0.42]	0.005 [0.37]	0.002 [0.15]	-0.014 [-0.99]	-0.004 [-0.26]	-0.013 [-0.91]	-0.001 [-0.09]
Average shoulder width (feet)	-0.088 [-1.79]	-0.106 [-3.11]a	-0.098 [-2.44]b	-0.092 [-2.39]b	-0.102 [-3.08]a	-0.117 [-3.54]a	-0.123 [-2.74]a	-0.123 [-5.19]a
Indicator for shoulder type curb	-0.819 [-1.52]	-0.97 [-2.53]a	-0.834 [-1.87]	-0.724 [-1.77]	-0.838 [-2.31]b	-1.088 [-2.91]a	-1.098 [-2.26]b	-1.154 [-4.34]a
Indicator for shoulder type wall	-1.002 [-1.79]	-1.216 [-3.38]a	-1.148 [-2.68]a	-1.105 [-2.65]a	-1.046 [-2.75]a	-1.133 [-3.07]a	-1.119 [-2.23]b	-1.313 [-5.08]a
AADT per lane	7.51E-05 [13.22]a	8.05E-05 [21.39]a	8.10E-05 [18.51]a	8.14E-05 [19.38]a	7.72E-05 [18.32]a	7.15E-05 [20.78]a	6.98E-05 [14.89]a	7.09E-05 [23.73]a
Speed limit (mile per hour)	-0.023 [-2.25]b	-0.011 [-1.28]	-0.014 [-1.43]	-0.015 [-1.50]	-0.018 [-2.39]b	-0.006 [-0.77]	3.75E-04 [0.06]	-0.019 [-2.82]b
Constant	-2.105 [-2.12]b	-2.066 [-2.63]a	-1.747 [-2.11]a	-1.39 [-1.63]	-0.333 [-.50]	-0.413 [-0.60]	-0.276 [-0.37]	-0.442 [-0.77]

*A randomizing procedure was added to the *qcount* package to make the estimation for lower quantiles independent of sequence of original data and the pseudo randomizing procedure of jittering sampling.
[a] Significant at 1%
[b] Significant at 5%

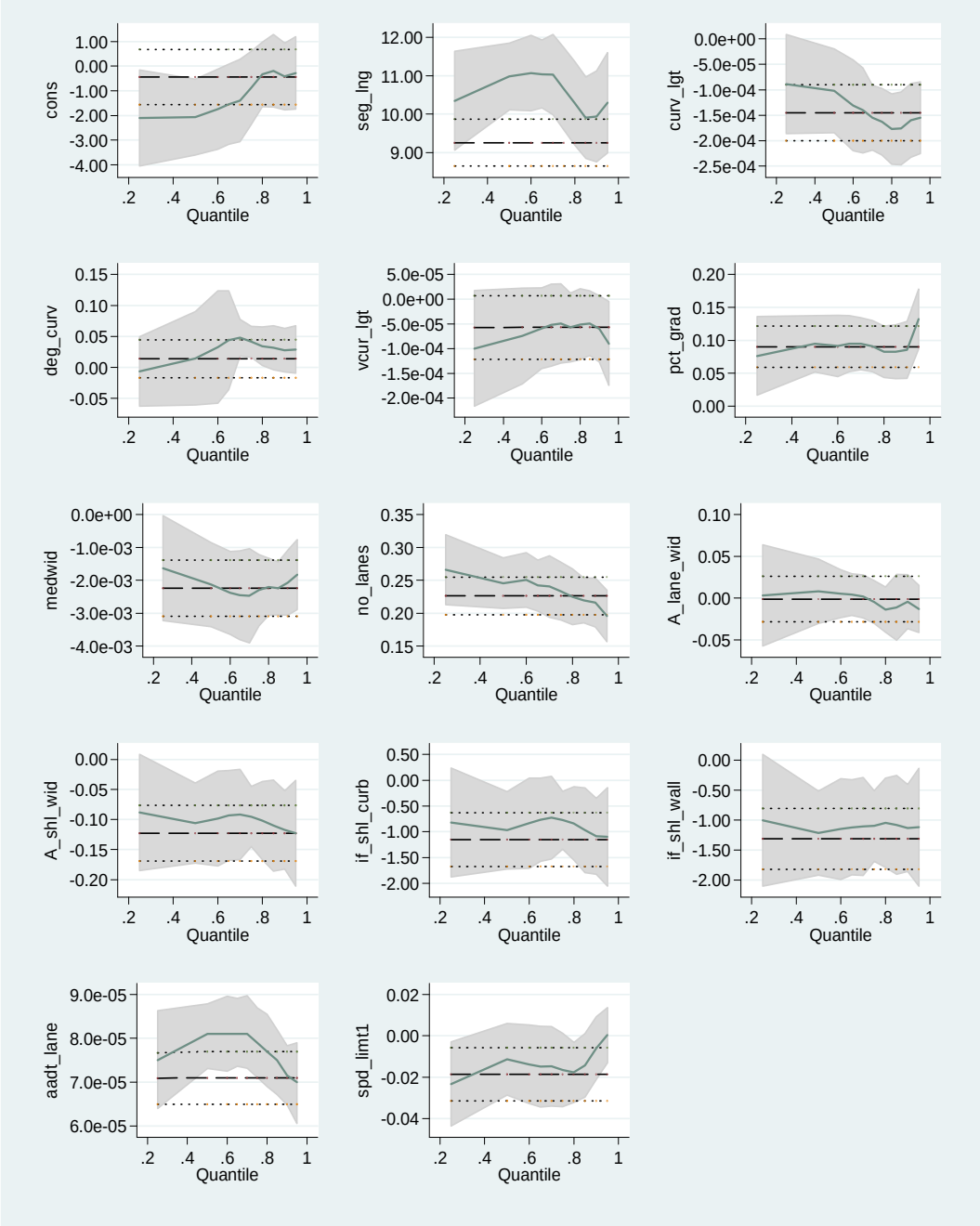


Figure 4.1: Parameter Estimates for QRs and NB Model for Urban Interstate Highway Segments

Estimates in Table 4.2 clearly show the statistical significance of the covariates and can be used directly for the purpose of prediction. However, the direct interpretation of coefficients in Table 4.2 may suggest some misleading conclusions, since $\boldsymbol{\beta}(\tau)$ are the linear partial effects on $Q_{T(z;\tau)}(\tau|\mathbf{x})$ but not on $Q_z(\tau|\mathbf{x})$ or $Q_{\mathrm{y}}(\tau|\mathbf{x})$. To fully understand the effects of covariates on the crash number, marginal effects of covariates on $Q_z(\tau|\mathbf{x})$ at the mean are presented in Table 4.3 and further illustrated in Figure 4.3.

Similar as defined in Figure 4.1, the dashed line in each figure in Figure 4.2 represents the NB regression estimate of the marginal effects. The two dotted lines show conventional 95 percent confidence intervals for NB regression estimate. The solid black line represents the quantile regression estimates of the marginal effects. The shaded gray area depicts a 95 percent confidence band for the quantile regression estimates.

All the marginal effects are calculated by setting the continuous variables to their mean and the dummy variables to zero. A few interesting findings from the estimation results are discussed as follows.

Table 4.3: Marginal Effects for QRs and NB Model for Urban Interstate Highway Segments

Variable	QR							Negative Binomial
	$Q_z(0.25\|x)$*	$Q_z(0.5\|x)$	$Q_z(0.6\|x)$	$Q_z(0.7\|x)$	$Q_z(0.8\|x)$	$Q_z(0.9\|x)$	$Q_z(0.95\|x)$	
Road segment length (miles)	4.171[a]	9.092[a]	11.710[a]	15.355[a]	20.238[a]	30.422[a]	42.897[a]	12.160[a]
Horizontal curve length (feet)	-3.74E-05	-8.88E-05	-1.47E-04	-2.32E-04	-3.80E-04	-5.28E-04	-6.95E-04	-1.91E-04
Degree of curvature (°/100ft)	-0.003	0.012	0.035	0.067[a]	0.067[b]	0.085	0.120	0.018
Vertical curve length (feet)	-4.30E-05	-6.49E-05	-6.48E-05	-7.05E-05	-1.03E-04	-1.91E-04	-3.98E-04	-7.52E-05
Vertical grade (%)	0.031[b]	0.078[a]	0.097[a]	0.132[a]	0.162[a]	0.262[a]	0.552[a]	0.119[a]
Median barrier width (feet)	-0.001[b]	-0.002[a]	-0.003[a]	-0.003[a]	-0.004[a]	-0.006[a]	-0.008[a]	-0.003[a]
Number of lanes	0.107[a]	0.203[a]	0.265[a]	0.335[a]	0.443[a]	0.661[a]	0.815[a]	0.297[a]
Average lane width (feet)	0.001	0.007	0.006	0.003	-0.027	-0.013	-0.054	-0.002
Average shoulder width (feet)	-0.073	-0.208[a]	-0.233[b]	-0.269[b]	-0.459[a]	-0.936[a]	-1.394[a]	-0.161[a]
Indicator for shoulder type curb	-0.235	-0.540[a]	-0.625	-0.744	-1.166[b]	-2.144[a]	-2.936[b]	-0.940[a]
Indicator for shoulder type wall	-0.278	-0.647[a]	-0.797[a]	-1.024[a]	-1.397[a]	-2.286[a]	-3.089[b]	-1.065[a]
AADT per lane	3.03E-05	6.66E-05	8.57E-05	1.13E-04	1.52E-04	2.19E-04	2.91E-04	9.31E-05
Speed limit (mile per hour)	-0.009[b]	-0.009	-0.015	-0.021	-0.035[b]	-0.018	0.002	-0.024[b]
Constant	4.171[a]	9.092[a]	11.710[a]	15.355[a]	20.238[a]	30.422[a]	42.897[a]	12.160[a]

*Marginal effects are calculated by setting the continuous variables to their mean and the dummy variables to 0.
[a] Significant at 1%
[b] Significant at 5%

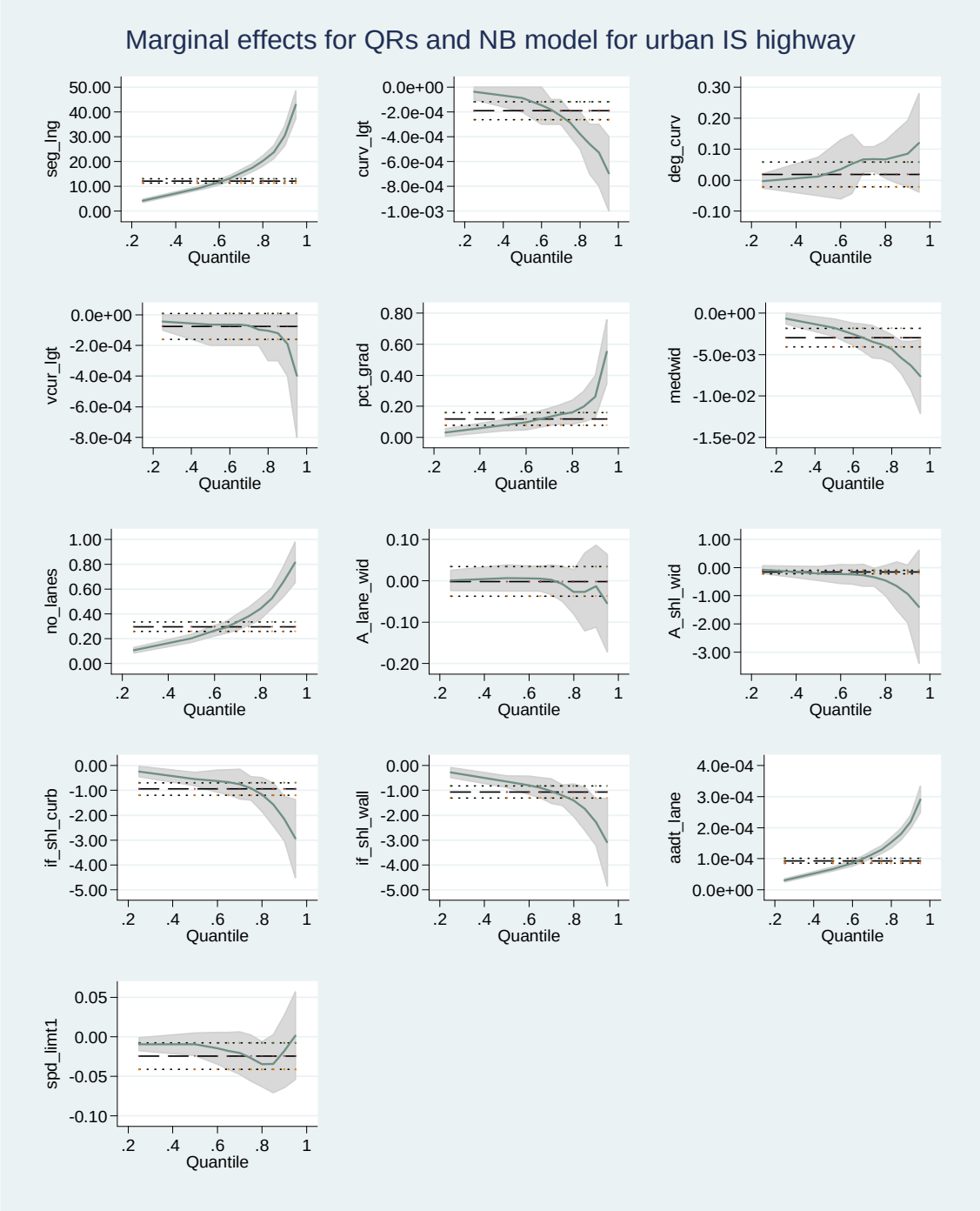


Figure 4.2: Marginal Effects for QRs and NB Model for Rural Interstate Highway segments

As shown in Table 4.3, it is apparent that QRs provides more information than the NB model. As expected, the road segment length has a large positive effect on the crash frequency. Holding all the covariates at their mean, the marginal effects on Q_z are much larger in the upper quantiles than those in the lower tail of the distribution. The size of the impacts at sixth decile is consistent with results obtained from the NB regression model. Moreover, the elasticities at each selected τ are calculated to be 0.38 (τ = 0.25), 0.41 (τ = 0.5), 0.42 (τ = 0.6), 0.44 (τ = 0.7), 0.44 (τ = 0.8), 0.46 (τ = 0.9) and 0.49 (τ = 0.95) respectively. It is interesting to notice that the elasticities are not equal to 1, indicating that if roadway segments are cut to smaller segments, different crash frequencies will be obtained. Similar results are found in related literatures (Anastasopoulos and Mannering, 2009) where it is explained as a "segment-boundary effects" on crash frequency in that changing geometrics may be associated with "crash clustering". This finding is one of the reasons that the crash number rather than crash rate is used as the response variable in this research.

For horizontal curves, their lengths have significant negative effects on Q_z across most part of the conditional distribution, which indicates that short curves and suddenly changes in curvature can increase the road risk. But the degree of curvature for horizontal curves is not found to be an important explanatory factor. This can be explained by the fact that most of the Interstate Highways are well designed and constructed so that the degree of the curvature should be qualified to offer a smooth and safe drive. For vertical curves, the length is insignificant at 1 percent across the conditional distribution. However, the vertical grade plays a significant role in affecting

road safety. Larger grades result in more dangerous sites and the marginal effects are higher in the upper tail of the distribution than those at other quantiles. This implies that with the same countermeasures, more crash reductions are expected for segments with more historical crashes given the same geometrical and traffic characteristics.

The number of lanes has a highly significant effect on Q_z at all selected quantiles, and the effects differ across the conditional distribution of crash data. This suggests that adding a lane to more dangerous sites will have more benefit in terms of road safety compared with those segments with equivalent characteristics. The marginal effect estimated by NB regression model is consistent with that of the QR model at about sixth decile. It is interesting to note that the marginal effects of average lane width on Q_z are surprisingly insignificant at all selected percentiles in QRs as well as in NB regression models. One possible explanation for this finding is that the dataset used in this study does not provide enough variation on average lane width, which has a mean of 12.721 ft and standard deviation of 2.515 ft. This is a reasonable case for Interstate Highways, where 12 ft to 14 ft is a typical design range.

A median barrier is presented for every segment in our dataset, and the width of a median barrier shows a significant negative effect on crash frequency, implying that wider median barriers reduce the crash frequency on average. Also, marginal effects differ across the conditional distribution, and NB regression model cannot reveal such detailed information.

Regarding roadway shoulder, both the type and average width have significant effects on Q_z in the QR and NB models. Results show that a wider shoulder reduces the

crash frequency more significantly. QRs show that risks derived from narrow shoulders are higher in those segments at the top of the conditional distribution than others. Shoulder type dummies have negative effects on Q_z. Presence of a wall or curb shoulder results in less traffic risks than presence of a level shoulder. As indicated in the methodology part, a covariate that affects Q_z does not guarantee to produce a marginal effect strong enough to change the conditional quantiles of the original crash counts, and one needs to be cautious with the interpretation of marginal effects on Q_y. The case for shoulder type gives a good example. Recall that $Q_y(\alpha|\mathbf{x}) = \lfloor Q_z(\alpha|\mathbf{x}) \rfloor$, marginal effects of the dummy indicating whether shoulder type is curb or not were calculated, and they are 0 (α = 0.25), -2 (α = 0.5), 0 (α = 0.6), -1(α = 0.7), -1 (α = 0.8), -2 (α = 0.9), and -3 (α = 0.95). At sixth decile, the marginal effect of the presence of a curb shoulder on the yearly crash counts is zero, although it has a significant effect on Q_z. At 95 percentile, the change from a level shoulder to a curb is enough to induce a marginal effect of +3 on the yearly crash counts.

In the group of covariates describing traffic characteristics, the effect of traffic volume is found to be highly significant. Similar to the findings for other covariates, it is found that the marginal effects are higher for segments in the higher tail than other parts of the distribution. The NB regression model only captures the mean effect of traffic volume on crash count which is around sixth decile. The elasticities at selected quantiles, 0.877 ($\alpha = 0.25$), 0.949 ($\alpha = 0.5$), 0.978 ($\alpha = 0.6$), 1.025 ($\alpha = 0.7$), 1.038 ($\alpha = 0.8$), 1.045 ($\alpha = 0.9$), and 1.075 ($\alpha = 0.95$), are all around 1. Unlike road segment length, AADT per lane tends to be a unit-elastic explanatory variable, which means that the

percentage change in Q_z caused by a percent change in traffic volume is equal to one and thus the crash rate is independent of traffic volume. Speed limit is found to be insignificant in this study; one possible reason is that the filtered dataset does not provide significant variations for the variable of speed limit.

4.3 Estimation of Crash Prediction Model for Rural IS Highways

In this section, a detailed discussion on crash prediction model for rural Interstate Highways is presented in the similar way as the discussion on urban Interstate Highways in section 4.2.

4.3.1 Data Description

The dataset used in developing the crash prediction model for Interstate Highway in rural areas contains 3,142 highway segments with an average length of 0.118 miles. The total number of reported crashes of all types is 2,042 and for each individual segment the crash rate ranges from 0 to 12. Compared to crash data for Interstate Highway segments in urban areas, the average number of crashes happened on Interstate Highway segments in rural areas in Washington State in 2002 were smaller. The average traffic volume on Interstate Highway in rural areas is 6,974, about one third of the average traffic volume on those in urban areas. The full descriptive statistics of the response variable and the covariates are presented in Table 4.4.

Table 4.4: Summary Statistics of Variables for Interstate Highways in Rural Areas in Washington State (3142 road segments)

Variable	Mean	Std. Dev.	Min	Max
Dependent variable				
Number of motor vehicle crashes (in 2002)	0.650	1.161	0	12
Covariates				
Geometric characteristics				
Road segment length (miles)	0.118	0.141	0.000227	1.73
Horizontal curve length (feet)	496.873	1034.163	0	6648
Degree of horizontal curvature (°/100ft)	0.448	0.911	0	6
Vertical curve length (feet)	766.899	680.993	0	4200
Vertical grade (%)	1.301	1.257	0	5.55
Median barrier width (feet)	98.565	142.012	4	999
Number of lanes	4.763	1.086	2	8
Average lane width (feet)	12.222	1.148	12	24.75
Average shoulder width (feet)	9.681	2.086	0	16
Indicator for shoulder type curb (1 if yes, 0 otherwise)	0.029	0.167	0	1
Indicator for shoulder type wall (1 if yes, 0 otherwise)	0.012	0.108	0	1
Traffic characteristics				
Average annual daily traffic (AADT) per lane	6974	3729	932	18475
Speed limit (mile/h)	68.919	3.567	35	70

4.3.2 Model Estimation and Results Analysis

The estimation was implemented using the *qcount* package of STATA on the same selected quantiles as discussed in section 4.2. Estimation results of the coefficients are presented in Table 4.5 and Figure 4.3. Estimation results of the marginal effects are presented in Table 4.6 and Figure 4.4. Because the marginal distribution of crash data for rural Interstate Highway segments shows that more than 50 percent of the segments have zero accident, the estimates of the quantile regression model tend to have large standard

errors for the quantiles below medium. It shows that extra zero observations in the distribution of dependent variable have an impact on the estimation accuracy and applicability of the proposed method.

Negative binomial regression model was also included and presented as the benchmark of comparison in this research. As shown in Table 4.6 and Figure 4.4, most of the covariates which are significant in the negative binomial regression model also tend to be significant in the quantile regression models. Among the geometric characteristics covariates, segment length, length and curvature of the horizontal curve, length and grade of vertical curve, number of lanes, and shoulder width are significant for most of the QRs and the NB model. For regressors that control traffic characteristics, the average annual daily traffic (AADT) per lane is significant under all the models, while the speed limit does not show any significant effect on crash frequency for this dataset. Moreover, the signs of the effects of the covariates that are significant do not switch at different quantiles in the study.

Compared to estimation results for Interstate Highway segments in urban areas, the significance for some covariates differs. The degree of curvature for horizontal curves, whose effect is highly insignificant for urban Interstate Highway segments, is significant in the NB and quantile regression models for rural Interstate Highway segments at 1 percent. Median barrier width for rural Interstate Highway segments is not significant, while its effect is highly significant for urban Interstate Highway segments. The effect of the shoulder type (curb/wall/ leveled shoulder) is insignificant for rural Interstate Highway segments, but significant for urban Interstate Highway segments.

Table 4.5: Parameter Estimates for QRs and NB Model for Rural Interstate Highway Segments (z-statistics in brackets)

Variable	QR							Negative Binomial
	$Q_z(0.25\|x)$*	$Q_z(0.5\|x)$	$Q_z(0.6\|x)$	$Q_z(0.7\|x)$	$Q_z(0.8\|x)$	$Q_z(0.9\|x)$	$Q_z(0.95\|x)$	
Road segment length (miles)	4.219 [10.45][a]	4.996 [19.53] [a]	5.491 [17.42] [a]	5.747 [16.22][a]	5.641 [16.57][a]	4.768 [10.91][a]	4.652 [12.51][a]	3.965 [20.83][a]
Horizontal curve length (feet)	-9.32E-05 [-1.65]	-7.71E-05 [-1.56]	-7.06E-05 [-1.63]	-8.71E-05 [-1.89]	-1.28E-04 [-2.27][b]	-8.66E-05 [-1.64]	-7.76E-05 [-1.99][b]	-9.53E-05 [-2.63][a]
Degree of curvature (°/100ft)	0.150 [2.54][b]	0.170 [3.45][a]	0.184 [3.93][a]	0.212 [4.09][a]	0.247 [4.34][a]	0.154 [3.37][a]	0.121 [3.27][a]	0.170 [5.19][a]
Vertical curve length (feet)	-1.53E-04 [-2.04][b]	-1.54E-04 [-2.34][b]	-1.61E-04 [-2.46][b]	-1.26E-04 [-1.82]	-8.79E-05 [-1.29]	-4.90E-05 [-0.8]	-7.22E-05 [-1.29]	-1.05E-04 [-2.33][b]
Vertical grade (%)	0.167 [4.36][a]	0.160 [4.61][a]	0.152 [4.43][a]	0.137 [3.73][a]	0.104 [2.98][a]	0.070 [2.3][b]	0.087 [2.75][a]	0.117 [4.64][a]
Median barrier width (feet)	6.44E-04 [2.09][b]	4.96E-04 [1.91]	4.33E-04 [1.92]	3.24E-04 [1.49]	2.91E-04 [1.19]	7.28E-05 [0.43]	-1.54E-04 [-0.73]	2.79E-04 [1.6]
Number of lanes	0.201 [4.84][a]	0.211 [5.35][a]	0.233 [6.18][a]	0.246 [6.5][a]	0.230 [6.65][a]	0.156 [5.25][a]	0.143 [4.94][a]	0.192 [7.26][a]
Average lane width (feet)	-0.446 [-0.08]	-0.369 [-0.1]	-0.325 [-0.6]	-0.323 [-2.09][b]	-0.353 [-2.26][b]	-0.242 [-1.69]	-0.175 [-1.82]	-0.250 [-2.66][a]
Average shoulder width (feet)	-3.52E-03 [-0.03]	-1.12E-04 [0]	-9.00E-03 [-0.18]	-2.35E-02 [-0.39]	-1.14E-02 [-0.17]	9.22E-02 [1.5]	8.07E-02 [0.86]	2.81E-02 [0.69]
Indicator for shoulder type curb	0.805 [0]	1.590 [0.32]	1.478 [0.83]	1.335 [1.49]	1.532 [1.56]	1.949 [2.28][b]	1.475 [1.48]	1.300 [2.17][b]
Indicator for shoulder type wall	-0.974 [0]	-0.114 [0]	0.180 [0]	-0.016 [0]	1.078 [0]	1.877 [1.4]	1.864 [1.47]	1.323 [1.84][b]
AADT per lane	1.02E-04 [8.53][a]	1.08E-04 [10.16][a]	1.08E-04 [10.62][a]	1.09E-04 [10.14][a]	1.14E-04 [10.02][a]	7.64E-05 [7.52][a]	6.27E-05 [6.49][a]	8.24E-05 [10.47][a]
Speed limit (mile per hour)	2.13E-02 [1.01]	1.19E-02 [0.91]	8.23E-03 [0.67]	6.84E-03 [0.54]	3.70E-03 [0.3]	-8.83E-03 [-0.88]	-7.22E-03 [-0.74]	5.52E-04 [0.06]
Constant	-0.303 [0]	-0.160 [0]	-0.286 [-0.04]	0.082 [0.04]	1.001 [0.46]	0.836 [0.45]	0.589 [0.38]	-0.078 [-0.06]

*A randomizing procedure was added to the *qcount* package to make the estimation for lower quantiles independent of sequence of original data and the pseudo randomizing procedure of jittering sampling.
[a] Significant at 1%
[b] Significant at 5%

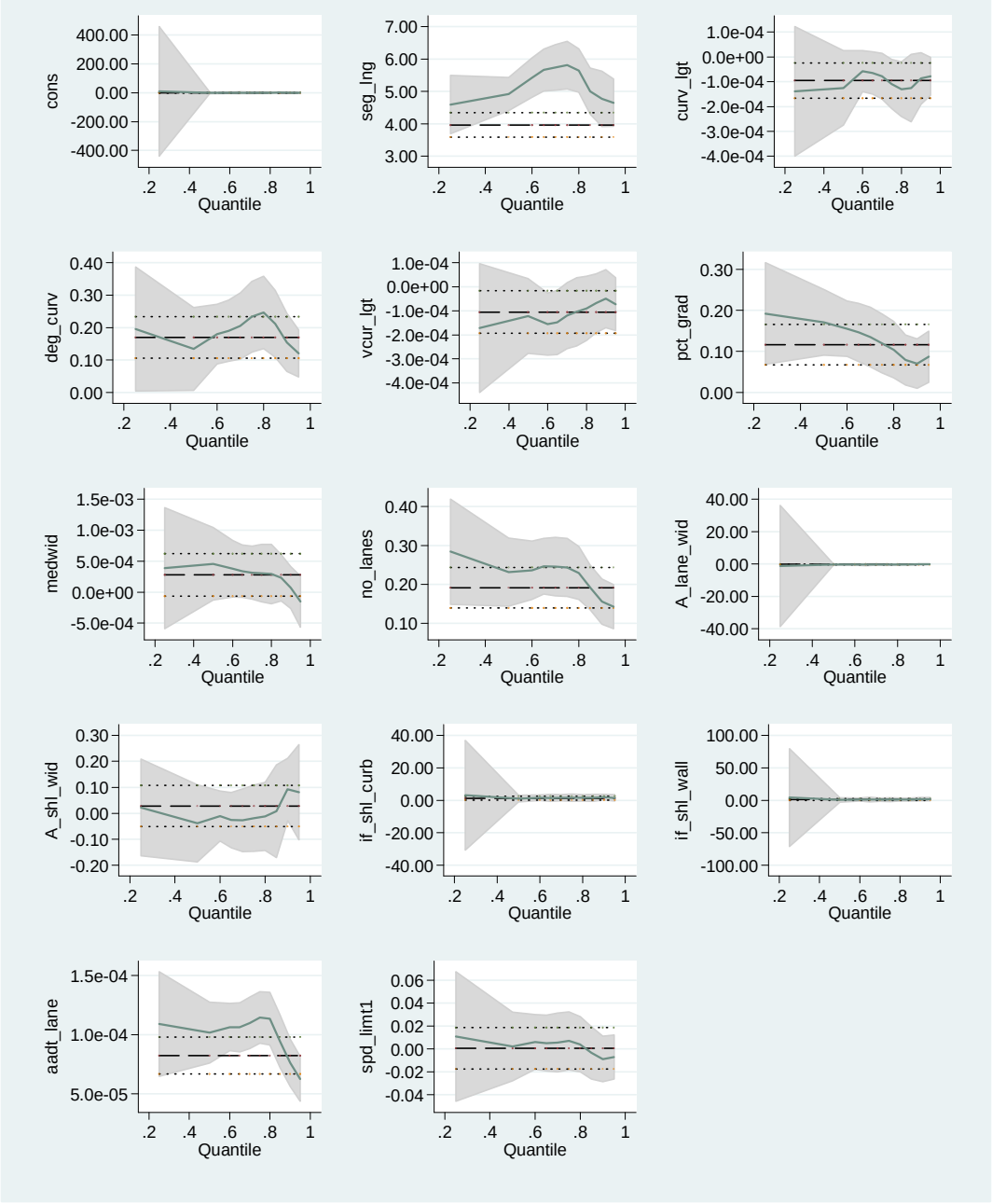


Figure 4.3: Parameter Estimates for QRs and NB Model for Urban Interstate Highway Segments

Table 4.6: Marginal Effects for QRs and NB Model for Rural Interstate Highway Segments

Variable	QR							Negative Binomial
	$Q_z(0.25\|x)$*	$Q_z(0.5\|x)$	$Q_z(0.6\|x)$	$Q_z(0.7\|x)$	$Q_z(0.8\|x)$	$Q_z(0.9\|x)$	$Q_z(0.95\|x)$	
Road segment length (miles)	0.503[a]	1.180[a]	1.613[a]	2.173[a]	3.146[a]	5.003[a]	7.102[a]	1.993[a]
Horizontal curve length (feet)	-1.18E-05	-1.92E-05	-2.17E-05	-3.49E-05	-7.80E-05[b]	-9.63E-05	-1.25E-04[b]	-4.79E-05[a]
Degree of curvature (°/100ft)	0.018[b]	0.040[a]	0.054[a]	0.080[a]	0.138[a]	0.162[a]	0.184[a]	0.086[a]
Vertical curve length (feet)	-1.83E-05[b]	-3.63E-05[b]	-4.74E-05[b]	-4.76E-05	-4.90E-05	-5.14E-05	-1.10E-04	-5.26E-05[b]
Vertical grade (%)	0.020[a]	0.038[a]	0.045[a]	0.052[a]	0.058[a]	0.074[b]	0.133[a]	0.059[a]
Median barrier width (feet)	7.19E-05[b]	1.11E-04	1.27E-04	1.23E-04	1.57E-04	7.58E-05	-2.35E-04	1.40E-04
Number of lanes	0.024[a]	0.050[a]	0.069[a]	0.093[a]	0.128[a]	0.163[a]	0.218[a]	0.096[a]
Average lane width (feet)	-0.053	-0.087	-0.095	-0.122	-0.197	-0.254	-0.267	-0.126
Average shoulder width (feet)	-4.34E-04	-2.65E-05	-2.64E-03	-8.87E-03	-7.04E-03	4.21E-02	1.23E-01	1.41E-02
Indicator for shoulder type curb	0.143	0.874	0.946	1.013	1.922	5.921[b]	4.898	1.293[b]
Indicator for shoulder type wall	-0.076	-0.025	0.058	-0.006	1.061	5.615	8.049	1.364[b]
AADT per lane	1.22E-05[a]	2.55E-05[a]	3.16E-05[a]	4.13E-05[a]	6.34E-05[a]	8.02E-05[a]	9.57E-05[a]	4.14E-05[a]
Speed limit (mile per hour)	2.54E-03	2.81E-03	2.42E-03	2.59E-03	2.07E-03	-9.27E-03	-1.10E-02	2.77E-04
Constant	0.503	1.180	1.613	2.173	3.146	5.003	7.102	1.993

Marginal effects are calculated by setting the continuous variables to their mean and the dummy variables to 0.
[a] Significant at 1%
[b] Significant at 5%

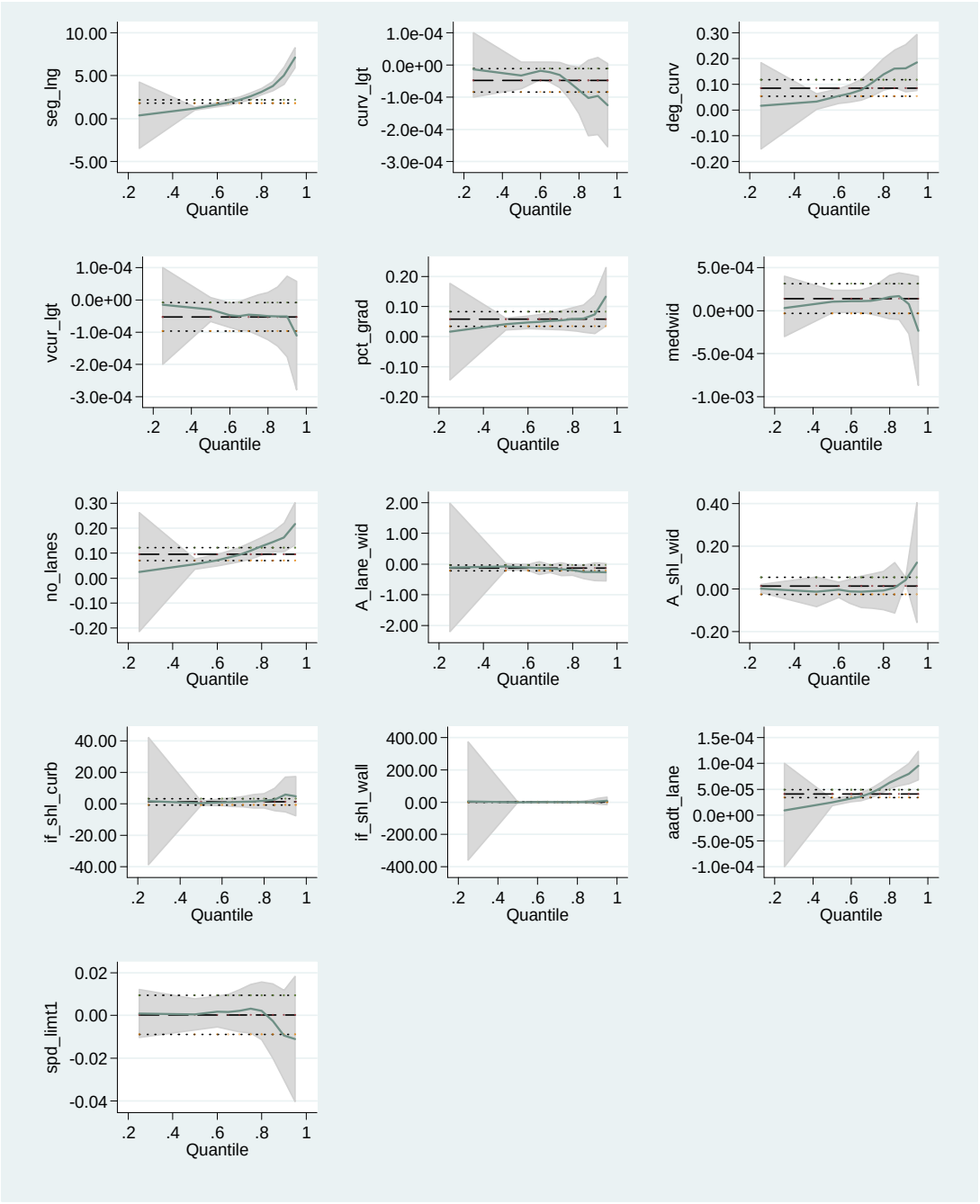


Figure 4.4: Marginal Effects for QRs and NB Model for Urban Interstate Highway Segments

As shown in the above tables and figures, QRs provides more information than the NB model. While some of the findings are consistent with those in Section 4.2, the others are different.

For horizontal curves, their lengths tend to have significant effects on the upper tail of the conditional distribution of Q_z, which indicates that short curves and suddenly changes in curvature can increase the road risk for risky highway sites. The degree of curvature has significant effects on Q_z across most part of the conditional distribution. For highway segments in the lower tail of the conditional distribution of Q_z, the estimated marginal effect is smaller than that estimated by NB model, while for those in the upper tail, the estimated marginal effect is larger than that estimated by NB model.

For vertical curves, the length is insignificant at 1 percent across the conditional distribution, but the vertical grade plays a significant role in affecting road safety. The size of the impacts at the seventh decile is consistent with results obtained from the NB regression model. Larger grades result in sites with larger risk and the marginal effects are higher in the upper tail of the distribution than those at other quantiles. This implies that with the same countermeasures, more crash reductions can be expected for segments with more historical crashes given the same geometrical and traffic characteristics.

Regarding roadway shoulder, the effects of both the type and average width are insignificant in the QR and NB models at 1 percent level for the rural segments, while both of them have significant effects for urban segments. It is noticed that highway segments in rural segments have wider and more leveled shoulders on average and significant reduction in traffic volume.

4.4 Model Validation and Comparison

After obtaining the parameter estimates for quantile regression models and NB model, data collected on I-82 was also used for model validation. For I-82, there are 1,213 homogeneous roadway segments with 1,014 segments in rural areas and 199 in urban areas. Both the Location method and the Probability method were applied to predict the number of crash for segments on I-82. For the Location method which requires the historical crash data as an input for the prediction, the average crash counts for the period of 1999 to 2001 were used as the historical crash data. The prediction results from the Location method, the Probability method, and the NB regression models were compared with the observed crash counts on I-82 in 2002; and the comparison results are summarized in Table 4.7.

There are various error measures in the estimation period, such as mean squared error, mean absolute error, mean absolute percentage error, and mean percentage error. The root of mean square error (RMSE) is a frequently-used simple measure of the differences between values predicted by a model or an estimator and the values actually observed from the variable being modeled or estimated. In this section, RMSE was adopted to measure the errors of prediction by different models for the following considerations. The RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{\sum_{i=1}^{\mathrm{N}} (\mathrm{Y}_i - \widehat{\mathrm{Y}}_i)^2}{\mathrm{N}}} \qquad (5.1)$$

where RMSE is the root of mean square error;

MSE is the mean square error;

N is the number of observations (road segments)

Y_i is the measured value of the i^{th} observation (the observed number of crashes in on segment I in the year 2002);

$\hat{Y}_i$ is the estimated value of the i^{th} observation (the predicted number of crashes in on segment I in the year 2002).

Table 4.7: Comparison of RMSEs for QRs and NB Regression Models

	RSME of Prediction		
	QR models		**NB**
	Probability Method	Location Method	
Overall	0.679	0.627	0.757
Rural IS Highway	0.688	0.621	0.771
Urban IS Highway	0.638	0.701	0.720

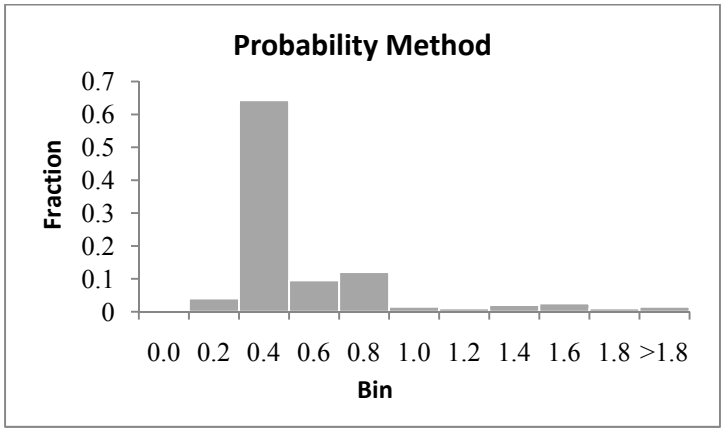
All RMSEs are showed in Table 4.7. The value indicates that on average the proposed QR based models have a better ability to predict crash occurrence than the NB regression model in this case study. Also, the RSMEs for Interstate Highway segments in urban areas are smaller than those in rural areas for the prediction by both the Probability Method and NB regression model. Such results are anticipated because there are significantly more segments with zero recorded crashes on the roadway segments in rural areas than those in urban areas, resulting in poor estimation. Furthermore, this table also shows that while the Location Method can generally produce better predictions than the Probability Method, it behaves worse than the Probability Method for Interstate

Highways in urban areas. This finding indicates that although the historical crash data can improve the prediction accuracy on average, it might not be the case under every situation. A traffic crash is a rare event which is largely subject to random fluctuations. The historical data used in this case study was collected over 3 years, implying that short-term recorded numbers and do not necessarily reflect the long-term expected numbers. Especially, since the number of roadway segments of I-82 in urban areas is relatively small, the impact of such fluctuation might become more obvious.

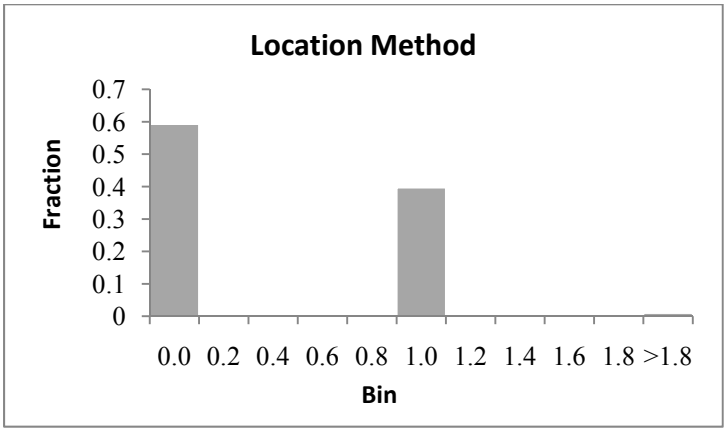
To compare the results in details, the histograms of the residuals which are the difference between the observed crash data on I-82 in 2002 and estimated values from either the proposed model or the NB regression model are present in Figure 4.5 and Figure 4.6. These histograms were normalized to showy relative frequencies. By using histograms, a visual impression of the distribution of the residuals by different models is clearly seen.

For predictions by the Probability Method and the NB regression model, historical crash data is not used. The distributions of the residuals by both methods are non-symmetrical and skewed to right. Compared to the NB regression model, the residuals from the Probability Method have smaller fractions distributed in the right tail.

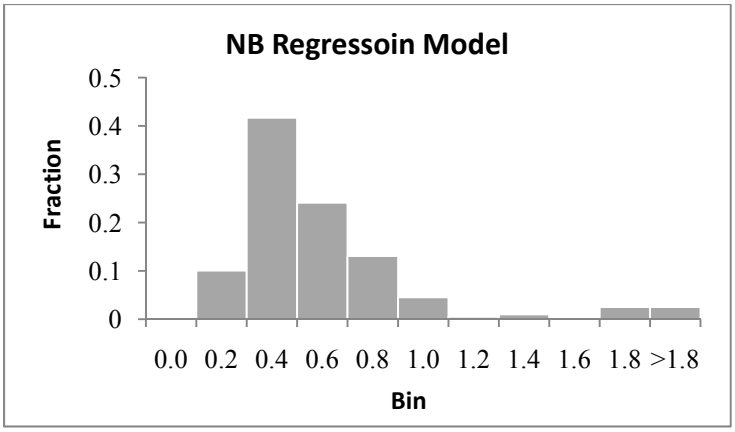
The location method takes the historical data into consideration, and provides discrete prediction values by its definition. In the histograms, it also shows that the Location Method precedes the NB regression model in the sense that is generates less large residuals.



(a) Probability Method

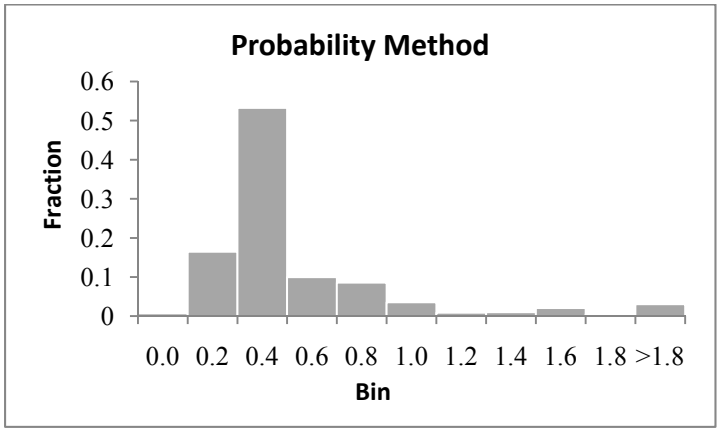


(b) Location Method

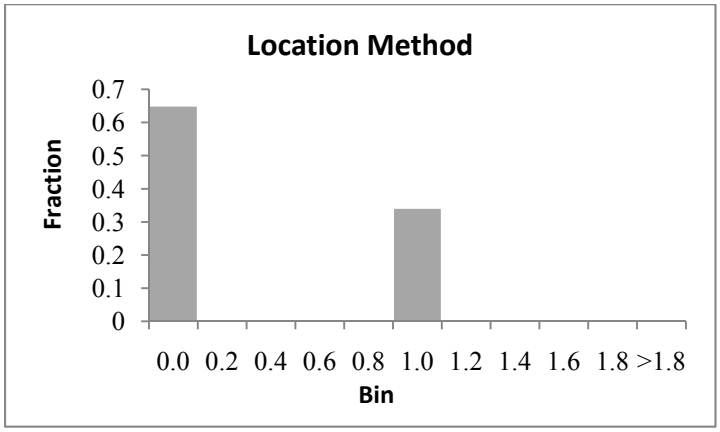


(c) NB Regression Model

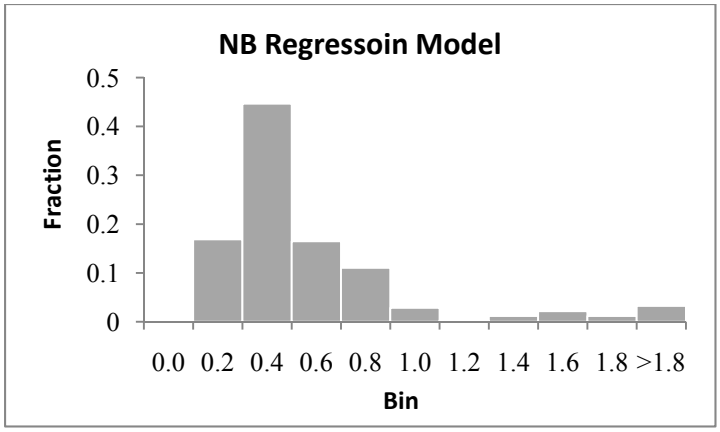
Figure 4.5: Histograms of the Residuals for Urban Interstate Highway Segments



(a) Probability Method



(b) Location Method



(c) NB Regression Model

Figure 4.6: Histograms of the Residuals for Rural Interstate Highway Segments

4.5 Summary

This chapter presented a case study for the purpose of demonstrating the application of the proposed crash prediction model, where the relation between crash occurrence and road inventory and traffic characteristics is investigated. The case study pertained to crash records and relevant HSIS data collected from Interstate Highways in both urban and rural areas in Washington State from 1999 to 2002. The results obtained by QRs were compared with those from the negative binomial regression. The results show that the conclusion on significance and signs of the effects derived from both the NB models the proposed QR models are generally consistent. Compared with the NB models, the proposed QR models reveal more detailed information on how the marginal effects of regressors on the crash rate change across the conditional distribution. Also, both the Probability Method and the Location Method as proposed prediction tools for the QR model provide better predictions compared to the prediction of NB regression models. Results from this chapter serve as the basis for developing the RRIs which are discussed in the next chapter.

CHAPTER 5 EMPIRICAL ANALYSIS ON DEVELOPMENT AND VISUALIZATON OF RRIS

As shown in Chapter 4, both the Location Method and the Probability Method can produce good prediction results in the selected dataset. In this chapter, the discussions are focused on using the results from the proposed prediction models to develop both the RRI_{Ind} and the RRI_{Acu} for the Interstate Highway 82 (I-82) based on the road geometry and traffic features observed in the year 2002. A Geographic Information System (GIS) platform is used to store, manage, and present the RRIs along with other related data using the linear referencing tools in ArcGIS. This delivery system supports all of the basic GIS functions, including viewing and querying the RRIs on the highway network, performing spatial analysis tasks such as selecting and buffering features, and manipulating information.

5.1 Data Description

The road geometry and traffic data collected on I-82 in Washington in the year 2002 was used to demonstrate the development of RRIs. For the Location methodwhich requires historical crash data as an input for the prediction, the average crash counts for the year 1999 to 2001 were used as the historical crash data. For I-82, there are 1,213 homogeneous road segments with 199 in urban areas and 1,014 segments in rural areas. The average lengths are 0.053 for segments in urban areas and 0.120 for those in rural

areas respectively. Vertical curves in rural areas have larger grade compare to those in urban areas on average. The average median barrier width for segments in rural areas is approximately 116 ft, which is about twice as much as that for segments in urban areas. The average traffic volume in rural areas (3,963 veh/day/lane) is significantly smaller compared to that in urban areas (8,346 veh/day/lane). The full descriptive statistics of the covariates are presented in Table 5.1 and Table 5.2.

Table 5.1: Summary Statistics of Variables for I-82 in Urban Areas in Washington State (199 road segments)

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Covariates				
Geometric characteristics				
Road segment length (miles)	0.053	0.055	0.000871	0.28
Horizontal curve length (feet)	259.292	516.502	0	1837
Degree of horizontal curvature (°/100ft)	0.405	0.932	0	5
Vertical curve length (feet)	834.673	995.319	0	3600
Vertical grade (%)	0.982	0.978	0	3
Median barrier width (feet)	58.613	26.659	40	200
Number of lanes	4.005	0.071	4	5
Average lane width (feet)	12.755	2.316	12	24.5
Average shoulder width (feet)	8.814	2.855	0	11
Indicator for shoulder type curb (1 if yes, 0 otherwise)	0.065	0.248	0	1
Indicator for shoulder type wall (1 if yes, 0 otherwise)	0.050	0.219	0	1
Traffic characteristics				
Average annual daily traffic (AADT) per lane	8346	2513	3506	10831
Speed limit (mile/h)	61.960	3.980	60	70

Table 5.2: Summary Statistics of Variables for I-82 in Rural Areas in Washington State (1014 road segments)

Variable	Mean	Std. Dev.	Min	Max
Covariates				
Geometric characteristics				
Road segment length (miles)	0.120	0.138	0.0001	1.14
Horizontal curve length (feet)	1209.973	1962.471	0	12683
Degree of horizontal curvature (°/100ft)	0.643	1.021	0	5.46
Vertical curve length (feet)	1515.301	1138.819	0	6700
Vertical grade (%)	1.775	1.622	0	5.01
Median barrier width (feet)	115.933	138.857	16	999
Number of lanes	3.982	0.132	3	4
Average lane width (feet)	12.452	1.622	12	23.5
Average shoulder width (feet)	9.418	2.156	0	11
Indicator for shoulder type curb (1 if yes, 0 otherwise)	0.035	0.183	0	1
Indicator for shoulder type wall (1 if yes, 0 otherwise)	0.024	0.152	0	1
Traffic characteristics				
Average annual daily traffic (AADT) per lane	3963	955	1663	6895
Speed limit (mile/h)	69.882	1.082	60	70

5.2 Estimation of RRIs

In Chapter 4, crash frequencies have been estimated by both the Location method and the Probability method, and the results were summarized in Table 4.7. In this section, the estimated crash frequencies by both methods were employed to develop both the RRI_{Ind} and the RRI_{Acu} for the Interstate highway 82 based on the road geometry, traffic features and historical crash data.

To estimate the RRIs, the predicted number of each type crash on each roadway segment was first calculated according to the distribution of the crash types for the highway group to which the segment belongs. The RRI_{Ind} and RRI_{Acu} were computed according to equation (3.2) and equation (3.4) discussed in Chapter 3. As discussed earlier, the estimated RRIs are sensitive to the selection of parameters for the risk sensitivity functions. The values of those parameters differ due to the different attitudes towards road risk among driver groups. In such a flexible formulation of the RRIs, the transportation agencies can adjust these parameters to accommodate the proposed RRIs to their local conditions. In this case study, the default value for these parameters was taken from Wu and Zhang (2008), where the parameters were determined by using a simulation dataset.

Table 5.3 summarizes the basic statistics of the estimated RRIs for the road segments in both rural and urban areas. As shown in the table, the average RRIs in rural areas are smaller than those in urban areas, which indicates that on average road segments of I-82 in rural areas are less risky compared to those in urban areas according to the definition of the RRIs.

Table 5.3: Summary Statistics of RRIs for Interstate Highway 82 in Washington

		I-82 in Urban Areas (199 Segments)		I-82 in Rural Areas (1,014 Segments)	
		Mean	Std. Dev.	Mean	Std. Dev.
RRI_{Ind}	Probability Method	4.0	3.3	2.9	3.2
	Location Method	4.0	4.5	3.0	3.9
RRI_{Acu}	Probability Method	4.6	3.8	2.3	3.1
	Location Method	4.2	4.5	2.7	3.7

While statistics in Table 5.3 give a general description of developed RRIs, maps can delivery such information in a more effective way. In the following section, ArcGIS, which is a popular and powerful tool to visualize data, was used to analyze data in relation to its location, leading to an increase in the value and utility of that information.

5.3 Visualizing RRIs using ArcGIS

The data source for the maps is the WSDOT GeoData Distribution Catalog *(*WSDOT, 2011*)*, a centralized distribution site for geographic information system (GIS) data produced at the Washington State Department of Transportation. The GIS data includes: lane information, Global Positioning System (GPS) route data, road log, etc. The State Route GPS Routes file is a collection of datasets representing Washington's State Routes created from "a combination of Global Positioning System data and inertial data are depicted as polylines with measures and State Route identifiers" (WSDOT, 2011). In this study, the State Route GPS Routes file was used to develop the base map for visualizing the estimated RRIs.

The Linear Referencing technique was applied to link the table containing the estimated RRIs, road geometry feature, and traffic features with the base map. There are quite a few software packages currently available to facilitate the implementation of Linear Referencing System, including the Linear Referencing tools in ArcGIS. Two primary data types are used to implement linear referencing in ArcGIS: route feature

classes which are a line feature class with a defined measurement system, and event tables which contain information about assets, conditions, and events located along route features.

At WSDOT, there are two basic types of the Distance Measuring Instrument (DMI) LRS (tabular) and Spatial LRS (graphical). The DMI LRS is created by driving the state highways with a vehicle mounted the Distance Measuring Instrument (DMI), a high accuracy odometer. The accuracy of DMI is about +/- 52.8 ft (WSDOT , 2009). A spatial LRS is based on the linear elements by X/Y coordinates in relationship to the earth's surface. WSDOT has three different spatial LRSs to increase the horizontal accuracy and level of details: the 500k LRS with accuracy of +/- 200 ft, the 24k LRS with accuracy of +/- 40 ft), and the GPS/LRS with accuracy of +/- 5 ft. In this research, the State Route GPS Routes datasets which implemented GPS/LRS were used as the source for the route feature classes for Linear Referencing. A state route in the State of Washington is identified by a three digit number (e.g., 082).

Distances along the linear feature are called measures and can be displayed in various units such as feet, miles, kilometers, or percentages. At WSDOT, road segment features are located along highways from their beginning to ending by Accumulated Route Mileage (ARM) and State Route MilePost (SRMP) values (WSDOT, 2009). Both ARM and SRMP values are carried to the hundredth of a mile. The SRMP is identified reference points and is usually used for the purpose of locating. The ARM is an accrual of mileage from the beginning of a route and is usually used for the purpose of calculating distance.

The event table was developed based on the RRIs derived from the previous section and road segment data abstracted from the HSIS dataset. For every road segment, its location on the base map is determined by three fields: the route identifier, beginning MilePost and ending MilePost. Roadway geometry features, traffic data and RRIs were included as the attributes for highway segment features in the table.

The geographical information of the seven interstate routes in Washington State is shown in Figure 5.1. I-82 was selected to demonstrate the development of the Road Risk Indices.

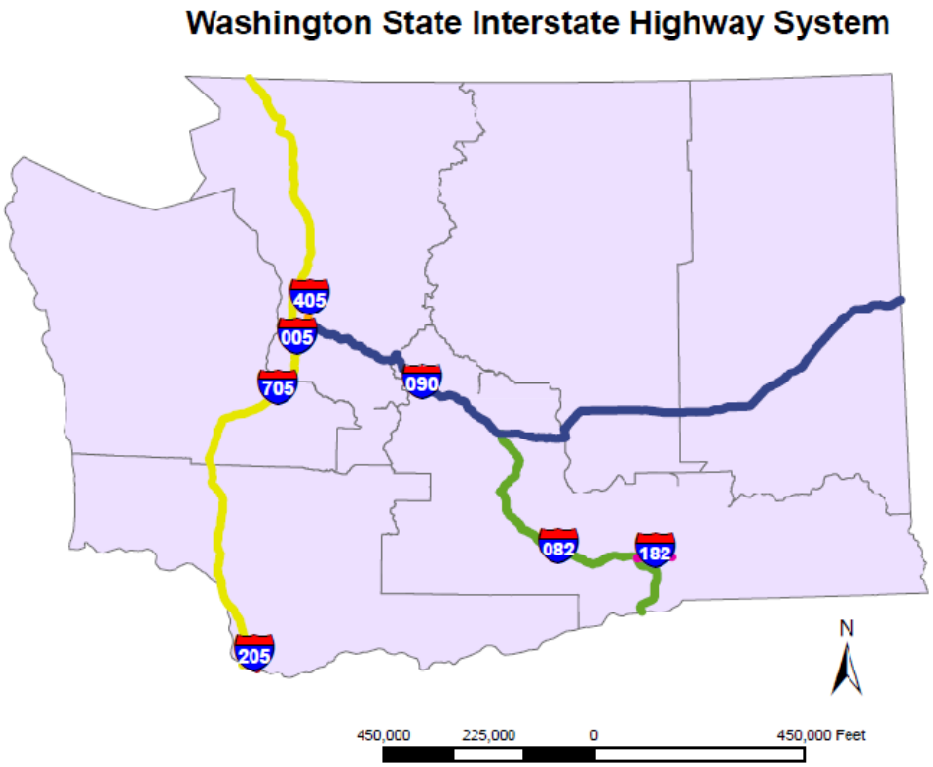


Figure 5.1: Map of the Washington State Interstate Routes

The risk maps developed according to the RRI_{Ind} and RRI_{Acu} by both the Probability Method and the Location Method are displayed in Figure 5.2 to Figure 5.5. Such maps enable comparison of the safety performance of interested sites, and make it easy for both road users and highway agencies to assess and grasp the safety information.

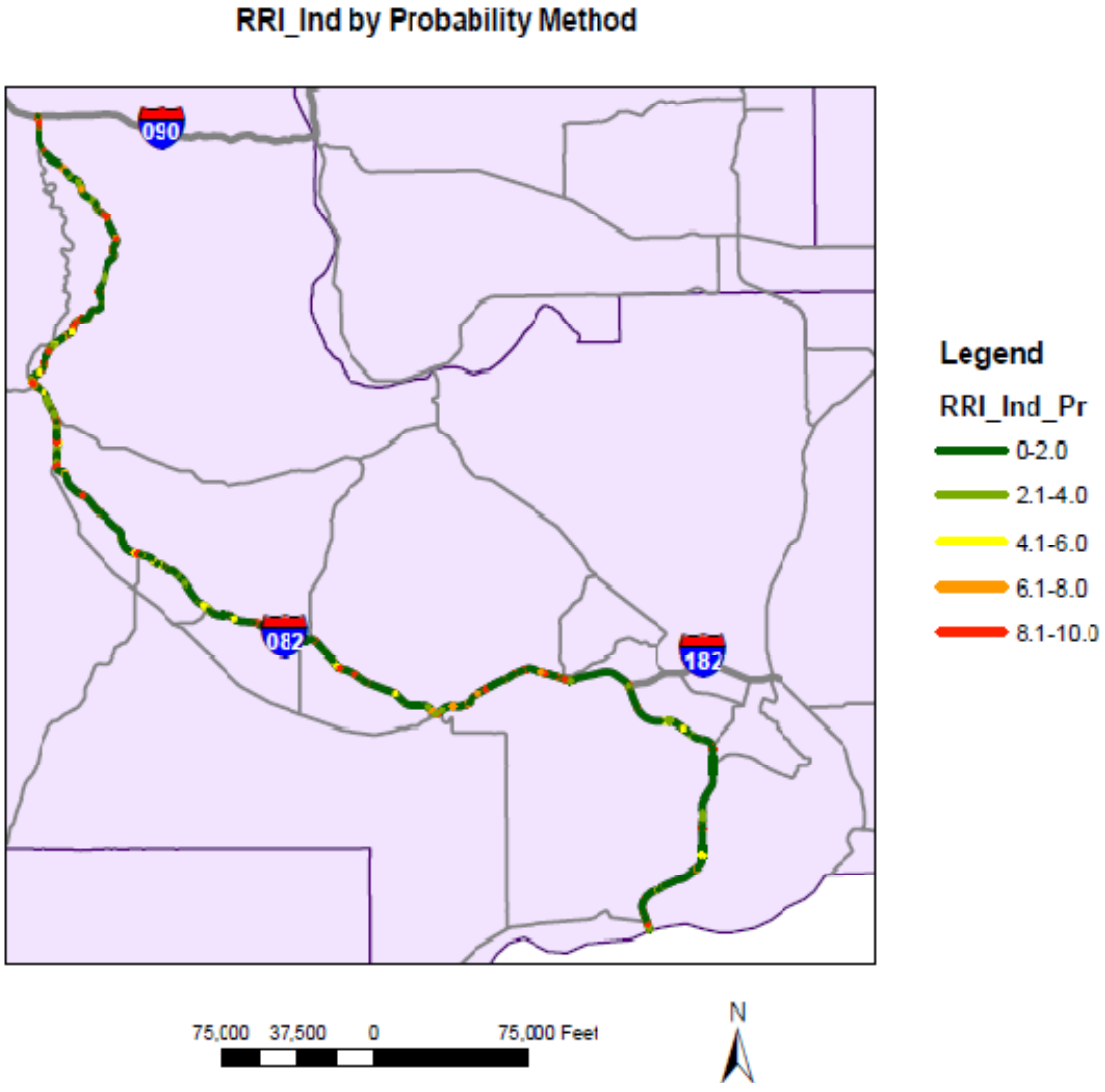


Figure 5.2: Map of RRI_{Ind} Estimated by the Probability Method

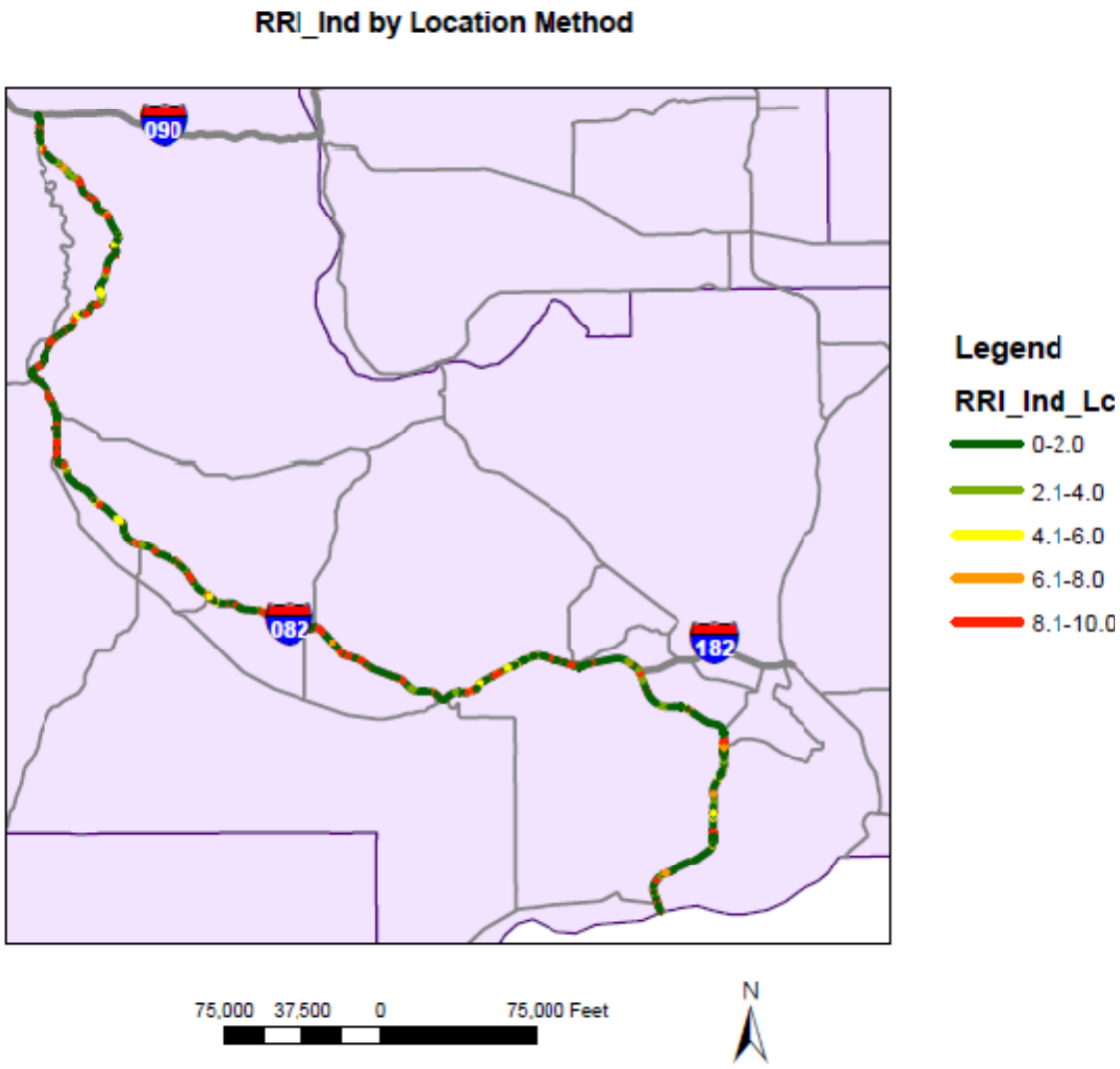


Figure 5.3: Map of RRI_{Ind} Estimated by the Location Method

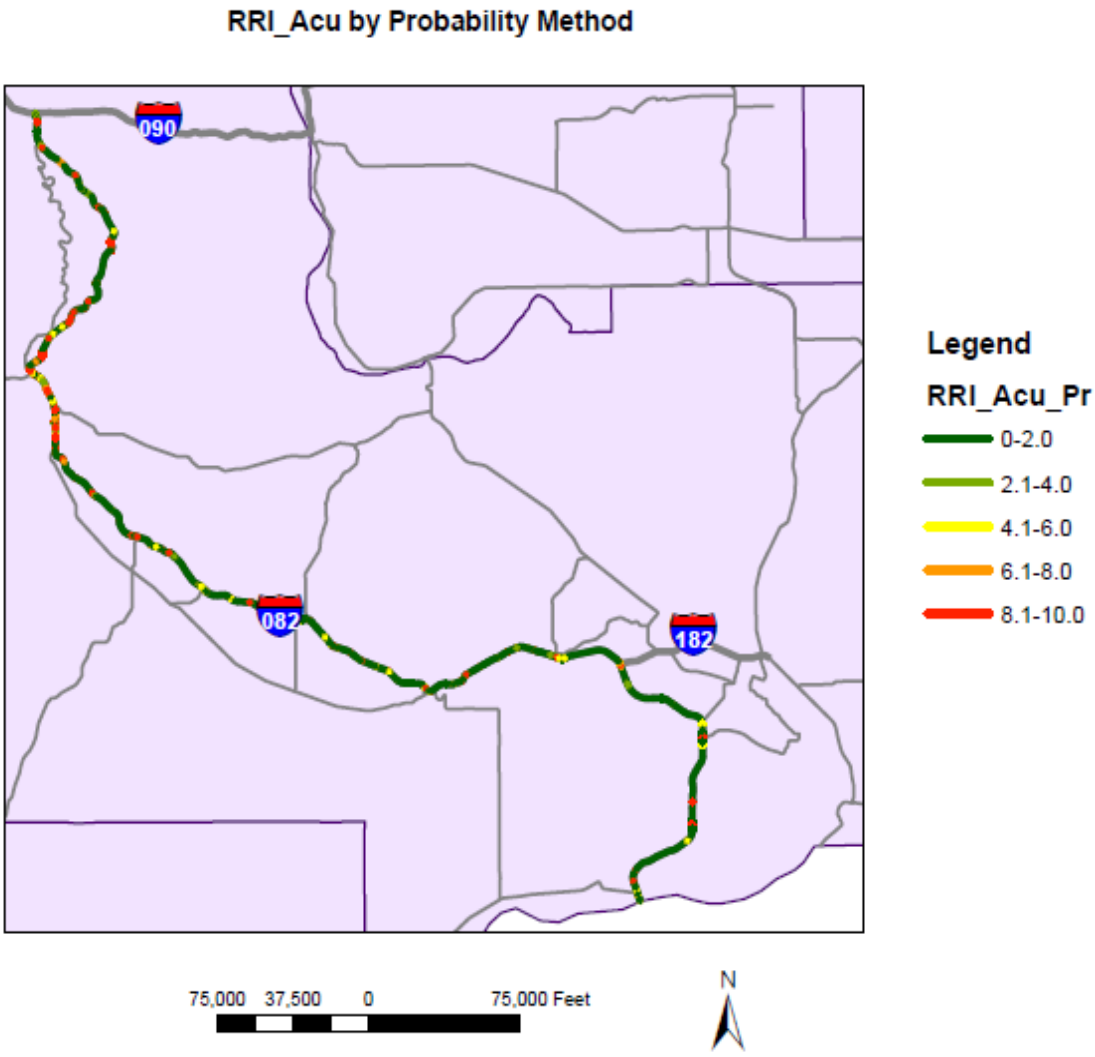


Figure 5.4: Map of RRI_{Acu} Estimated by the Probability Method

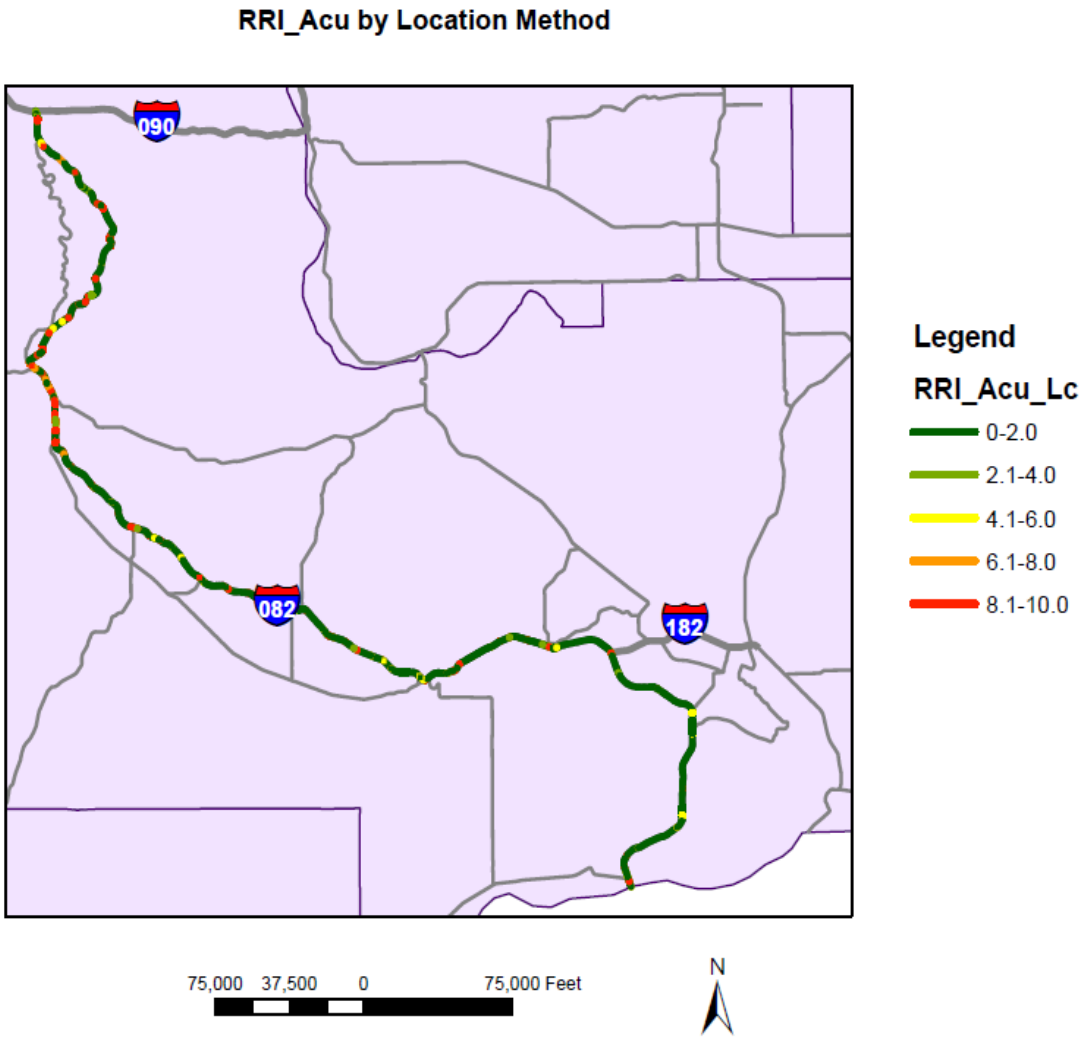


Figure 5.5: Map of RRI_{Acu} Estimated by the Location Method

5.5 Summary

This chapter presents a case study on developing both the RRI_{Ind} and the RRI_{Acu} using the calibrated crash prediction models from the previous chapter. A Geographic Information System (GIS) platform is developed to store, manage, and

present RRIs and related data with reference to geographic location data using the linear referencing tools in ArcGIS. The results of the case study show that the proposed methodology framework is capable of assessing and effectively delivering information on road risks of existing highway infrastructure for both road users and highway agencies based on road geometrics, traffic conditions, and historical crash data. In the final chapter of this dissertation, the major findings and topics for future research are presented.

CHAPTER 6 SUMMARY AND RECOMMENDATIONS

This dissertation research addressed an important highway safety issue with the development of a methodological framework for assessing road safety performance of existing highway infrastructures through composite indices. This chapter highlights the main findings from this research and provides recommendations for using the proposed methodology in highway safety research and practice. The chapter ends with a discussion on possible topics with which the research can be extended in the future.

6.1 Summary of Research Findings

In keeping with the original goals and objectives established for the dissertation research, key findings from this research work include:

1. In order to assess road safety and further facilitate the safety management of the existing highway networks in a proactive and objective manner, a set of composite indices that can quantify the potential road risks is needed. These indicators are appealing for at least three reasons. First, the indices can evaluate the potential road risk according to road geometry features, traffic features, and historical crash data provided that such data is available. Considering the fact that traffic crashes are complex and rare events, such indices can provide a more reliable and proactive assessment of road risk

compared to an evaluation which is purely based on historical crash data. Second, these indices rely less on safety engineer's subjective experience and judgment and more on objective data and information. Finally, such composite indices aggregate various kinds of information into composite indicators. Compared to the current case-by-case assessment process in the U.S., the proposed indices can aid the safety management activities such as optimization and project prioritization at a network level, making it possible for safety management to be integrated in the process of highway infrastructure management.

2. A comprehensive framework for developing the Road Risk Indices (RRIs) is proposed in this dissertation. These RRIs are capable of assessing the road risks of existing highway infrastructure from the perspective of both the road users and the highway agencies. The proposed framework is easy to adopt state-of-the-art crash prediction models and provides the flexibility to accommodate any available historical data. Results from the case study show that the proposed RRIs are capable of integrating the results from the crash prediction models and the historical crash data in forming more general indices for road risk assessment.
3. A model using quantile regression for counts techniques is proposed as a methodological alternative to modeling traffic crash occurrences. Over-dispersion caused by unobserved heterogeneity is common in traffic crash data and failure to accommodate such heterogeneity in the modeling process

can lead to undermine the validity of the models. The Negative Binomial (NB) regression models have been adopted widely for accommodating over-dispersion in highway safety study. However, previous studies conducted by other researchers indicate that empirical crash data do not follow the underlying distribution assumptions of NB regression models. In this dissertation, a crash prediction model based on quantile regression for counts techniques is proposed to overcome some of the major limitations (including over-dispersion) of the traditional count models. The proposed model is designed as a semiparametric model which allows researchers to relax restrictions in the form of the distribution function of the response variable, resulting in more robust estimation. Also, the quantile regression based estimation technique estimates various quantiles of the conditional distribution of parameters rather than just the mean of a population, and thus provides a more complete picture of effects of covariates on crash frequency compared to traditional models. In addition, the proposed model can provide better predictions by effectively addressing the heterogeneity of crash data. A case study was carried out to demonstrate how to apply the proposed model to the crash records and relevant HSIS data collected on Interstate Highways in both urban and rural areas in Washington State in the year 2002. The results from the numerical case study show that the proposed model can provide a more complete analysis of crash data and yield more accurate predictions compared to the NB regression model.

4. Extra zero observations in crash counts are found to be a major influencing factor for prediction accuracy. In the numerical case study, separate analyses were carried out for the Interstate Highway segments in urban and rural areas considering the difference in the data collection process and the quality of crash data. The comparison of the results indicates that both the proposed model and the NB regression model behave better on the dataset for urban Interstate Highway segments where less zeros are observed. While the extra zeros have an impact on all the estimates of the NB regression model, they only influence the estimates in lower quantiles of the proposed model. For the estimates of the conditional quantiles on the upper tail of the distribution, which are also the focus of highway safety study, extra zeros do not affect the estimation accuracy.
5. The Geographic Information System (GIS) and linear referencing techniques can be applied to developing an effective and user friendly publishing system for the Road Risk Indices and related data with reference to their geographic locations. Safety data usually are collected from various procedures and stored in different formats, most of which are described as either point or linear features. The current GIS and linear referencing techniques enable researchers and users to combine the information from different resources, and store, manage, and present them in an integrated system. The numerical case study shows that ArcGIS and its Linear Referencing tool package provide an easy environment to fulfill the proposed GIS delivering system.

6.2 Topics for Future Research

This dissertation undertook a significant amount of work in establishing a methodological framework for developing the Road Risk Indices to assess road safety of existing highway networks. While the objectives of this research have been achieved, there are some valuable extensions that merit future research.

1. In this research, while the discussions and the numerical case study are focused on Interstate Highway segments, the proposed methodological framework can be customized for the safety study of other function classes of highways as well as intersections to form a more comprehensive evaluation of a highway network. For intersections, both the modeling procedure and the covariates of the crash prediction models may differ from the model developed in this dissertation, but the concepts and techniques proposed in this dissertation can be easily extended for such cases.
2. In the case study, simulated data is used for the parameters in the risk sensitivity functions. While research on determining those parameters for road safety study is not currently available, studies on similar topics in other areas have been found and techniques are available for fulfilling this task. A valuable contribution can be made by using such data to calibrate those parameters.

3. Estimating the models and developing the GIS mapping systems requires coding work and using different software under various development

environments. To make estimating and updating models easier, it is recommended that a software package be developed to consolidate all components of the proposed methodological framework into a single platform.

4. This study is based on cross section data; future work should be undertaken to extend the methodology for panel data. In addition, independent variables that change over time and have a potential impact on road risk levels should be considered. For example, pavements experience wear and deterioration from vehicle loading, resulting in lower coefficient of friction over time. Because the friction developed between a vehicle's tires and the pavement surface plays a key role in allowing drivers to exercise more control of their vehicles, deteriorated surface skid resistance could result in an increasing in crash rate and road risk levels. It would be interesting to consider such a variable in the model and explore its impacts on road safety. A model based on panel data could provide a general picture of how the road safety performance changes over time.

BIBLIOGRAPHY

Abdel-Aty, M.A., and Pemmanaboina, R. (2005). Assessing Crash Occurrence on Urban Freeways using Static and Dynamic Factors. *Advances in Transportation Studies, An International Journal,*5, pp. 39-51.

Adams, T.M.,Koncz, N.A., and Vonderohe A.P.(1997).*Guidelines for the Implementation of Multimodal Transportation Location Referencing Systems.* NCHRP Report 460. National Cooperation Highway Research Program, Washington, D.C. Retrieved online at http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_460.pdf.

Anastasopoulos P.C., and Mannering F.L. (2009). A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accident Analysis and Prevention* 41(1), pp.153–159.

ASSHTO (2010). Highway Safety Manual Introduction, retrieved online at http://www.highwaysafetymanual.org/.

ASSHTO, 2001. A Policy on Geometric Design of Highway and Streets. American Association of State Highway and Transportation Officials, Washington, D.C.

Brännäs, K., and Rosenqvist, G. (1994). Semiparametric Estimation of Heterogeneous Count Data Models. *European Journal of Operational Research* 76 (2), pp. 247-258.

Cade, B.S., and Noon, B.R. (2003). A Gentle Introduction to Quantile Regression for Ecologists. *Frontiers in Ecology and the Environment* 1(8), pp.412-420.

Cameron, A. C., and Trivedi, P. K. (1998). *Regression Analysis of Count Data,* Cambridge, U.K.: Cambridge University Press.

Council, F.M., and Mohamedshah, Y. (2009). *The Highway Safety Information System guidebook for the Washington state data files. Volume I: SAS File Formats.* Report No. , FHWA-RD-95-206. Federal Highway Administration, Washington, DC.

Daniel, J., and Chien, S.I.J. (2004). Truck Safety Factors on Urban Arterials. *Journal of Transportation Engineering* 130(6), pp.742-752.

Elvik, R., and Vaa,T. (2004). *The Handbook of Road Safety Measures*, Elsevier Science, Oxford.

Elvik, R. (2007). *State-of-the-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks*. Report 883, Institute of Transport Economics, Oslo, Norway.

Elvik, R., 2008. Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads. *Transportation Research Record 2083*, 72-75.

El-Basyouny, K., and Sayed, T. (2006). Comparison of two negative binomial regression techniques in developing accident prediction models. *Transportation Research Record 1950*, pp. 9-16

EuroRAP (2006a). What is EuroRAP? Retrieved online at http://www.eurorap.org/what_is_eurorap.

EuroRAP (2006b). What is Star Rating? Retrieved online at: http://www.eurorap.org/rps.

EuroRAP (2006c). What is Risk Mapping? Retrieved online at http://www.eurorap.org/riskmap.

Esri (2009). ArcGIS 9.2: An Overview of Linear Referencing Accessed online at http://webhelp.esri.com/arcgiSDEsktop/9.3/index.cfm?TopicName=An_overview_of_linear_referencing.

Friedman, M., and Savage, L. J. (1948). The Utility Analysis of Choices Involving Risk. *Journal of Political Economy* 54(4), pp.279-304.

FHWA (2009). Road Safety Audits (RSA), retrieved online at: http://safety.fhwa.dot.gov/.

Golias, J., Matsoukis, E., Yannis, G. (1997). An Analysis of Factors Affecting Road Safety: The Greek Experience. *Journal of the Institute of Transportation Engineers* 67(11), pp. 26-31.

Greene, W. (2008). *Functional Forms for the Negative Binomial Model for Count Data.* Economics Letters 99 (3), pp. 585-590.

Guikema, S.D., and Coffelt, J.P. (2008). A Flexible Count Data Regression Model for Risk Analysis. *Risk Analysis* 28 (1), pp. 213 - 223.

Hakkert, A.S, Gitelman, V. and Vis, M.A. (Eds.) (2007*) Road Safety Performance Indicators: Theory*. Deliverable D3.6 of the EU FP6 project SafetyNet.

Hall, R.W. (1996). Route Choice and Advanced Traveler Information Systems on a Capacitated and Dynamic Network. *Transportation Research C* 4(5), pp.289-306.

Hauer, E. (1997). *Observational Before-after Studies in Road Safety*. Pergamon Press, Elsevier Science Ltd., Oxford, UK.

Hauer, E. (2001). Overdispersion in Modeling Accidents on Road Sections and in Empirical Bayes Estimation. *Accident Analysis and Prevention* 33(6), pp.799-808.

Hauer, E., Kononov, J., Allery, B., and Griffith, M.S. (2002). Screening the Road Network for Sites with Promise. *Transportation Research Record 1784*, pp.27-32.

Hilbe, J.M. (2007). *Negative Binomial Regression*. Cambridge University Press, UK.

Harwood, D.H., Council, F.M., Hauer, E., Hughes, W.E., Vogt, A. (2000*). Prediction of the Expected Safety Performances of the Rural Two-Lane Highways.* In Report FHWA-RD-99-207, Washington, D.C.

Ivan, J.N., and Wang, C., (2000). Explaining Two-Lane Highway Crash Rates using Land Use and Hourly Exposure. *Accident Analysis and Prevention* 32(6), pp.487-795.

Joshua, S.C., and Garber, N.J. (1990). Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation Planning and Technology* 15(1), pp.41-58.

Kanellaidis, G., Yannis, G., and Harvatis, M. (1999). Attitude of Greek Drivers towards Road Safety. *Transportation Quarterly* 53(1), pp. 109-121.

Koenker, Roger W. and Gilbert Basset Jr. (1978) Regression Quantiles, *Econometrica* 46(1), pp. 33-50.

Koenker, R. and Park, B. (1996) An Interior Point Algorithm for Nonlinear Quantile Regression, *Journal of Econometrics* 71(1-2), pp. 265-283.

Koenker, R. (2004). Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis* 91(1), pp.74–89.

Koenker R. (2005) *Quantile Regression*, Cambridge University Press, New York.

Kumara, S.P., and Chin, H.C. (2003). Modeling Accident Occurrence at Signalized T Intersections with Special Emphasis on Excess Zeros. *Traffic Injury Prevention*, 4(1), pp.53-57.

Kumara, S.P., and Chin, H.C. (2005). Application of Poisson Underreporting Model to Examine Crash Frequencies at Signalized Three-Legged Intersections. *Transportation Research Record 1908*, pp.46-53.

Land K.C., McCall, P.L., and Nagi, D.S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models. *Sociological Methods and Research* 24 (4), pp.387-442.

Li, X., Lord, D., Zhang, Y., and Xie, Y., (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis and Prevention* 40 (4), pp. 1611-1618.

Lord, D., (2000). *The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Application of Accident Prediction Models*. Ph.D. Dissertation, Department of Civil Engineering, University of Toronto, Toronto, Ontario.

Lord, D., Washington, S.P., and Ivan, J.N. (2005). Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis and Prevention* 37(1), pp.35-46.

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I.,Ivan, J.N., Lyon, C., and Jonsson, T. (2009). Methodology for Estimating the Safety Performance of Multilane Rural Highways. NCHRP Web-Only Document 126,

National Cooperation Highway Research Program, Washington, D.C. Retrieved online at http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w126.pdf.

Lord, D., Washington, S.P., and Ivan J.N. (2007). Further notes on the application of zero inflated models in highway safety. *Accident Analysis and Prevention* 39 (1), pp. 53-57.

Leur D.P., and Sayed, T., (2002). Development of a Road Safety Risk Index, *Transportation Research Record 1784*, pp. 33–42.

Lynam, D., Hummel, T., Barker, J. and Lawson, S. (2004). *European Road Assessment Programme 1.* Accessed online at http://www.eurorap.org.

Ma, L., and Pohlman, L. (2008). Return Forecasts and Optimal Portfolio Construction: A Quantile Regression Approach. *The European Journal of Finance* 14(5), pp.409-425.

Ma, J. (2006). Bayesian Multivariate Poisson-Lognormal Regression for Crash Prediction on Rural Two-Lane Highways. Ph.D. Dissertation, the University of Texas at Austin.

Machado, J.A.F., and Santos Silva J.M.C.(2005). Quantiles for Counts. *Journal of the American Statistical Association* 100(427), pp. 1226-1237.

Miaou, S.P., Hu, P.S., Wright, T., Rathi, A.K., and Davis, S.C. (1992). Relationship between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach. *Transportation Research Record 1376*, pp.10-18.

Miaou, S.P., and Lum, H. (1993). Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis and Prevention* 25(6), pp.689-709.

Miaou, S.P. (2001). *Estimating Roadside Encroachment Rates with the Combined Strengths of Accident- and Encroachment-Based Approaches.* Publication FHWA-RD-01-124. FHWA, U.S. DOT.

Miaou, S.P., and Lord, D. (2003). Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes. *Transportation Research Record 1840*, pp.31-40.

Mitra, S., and Washington, S. (2007). On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. *Accident Analysis and Prevention* 39 (3), pp.459–468.

Miranda, A. (2006). QCOUNT: Stata program to fit quantile regression models for count data, Statistical Software Components S456714, Boston College Department of Economics.

Miranda, A. (2008). Planned Fertility and Family Background: A Quantile Regression for Counts Analysis. *Journal of Population Economics* 21(1), pp.67-81.

Montella, A. (2005). Safety Reviews of Existing Roads: a Quantitative Safety Assessment Methodology. *Transportation Research Record 1922*, pp.62-72.

Moreira, S. and Pita Barros, P. (2009). Double Coverage and Demand for Health Care: Evidence from Quantile Regression. *Health, Econometrics and Data Group (HEDG) Working Papers, HEDG,* c/o Department of Economics, University of York.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A. and Giovannini, E. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*, Paris: OECD.

NHTSA (2009). *Traffic Safety Facts 2009*. National Highway Traffic Safety Administration, Report DOT HS 811 140. Washington, D.C.

Park, B.-J., and Lord, D. (2008). Adjustment for the Maximum Likelihood Estimate of the Negative Binomial Dispersion Parameter. *Transportation Research Record* 2061, pp.9-19.

Park B., and Lord, D. (2009). Application of Finite Mixture Models for Vehicle Crash Data Analysis. *Accident Analysis and Prevention* 41(4), pp.683-691

Park, B. (2010). *Application of Finite Mixture Models for Vehicle Crash Data Analysis.* . Ph.D. Dissertation, Department of Civil Engineering, Texas A&M University.

Paul, D.L., and Tarek ,S. (2002). Development of a Road Safety Risk Index. *Transportation Research Record 1784*, pp.33-42.

Shankar, V.N., Mannering, F., and Barfield, W. (1995). Effect of Roadway Geometric and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention* 27 (3), pp.371–389.

Shankar, V., Milton, J., and Mannering, F.L. (1997). Modeling Accident Frequency as Zero Altered Probability Processes: An Empirical Inquiry. *Accident Analysis and Prevention* 29(6), pp.829-837.

Stephens, D. (1990). Risk and Incomplete Information in Behavioral Ecology. In: E. Cashdan (Ed.), *Risk and Uncertainty in Tribal and Peasant Economics.* Boulder, CO: Westview Press. pp. 19–46.

Transfund New Zealand (2003). Safety Audits of Existing Roads: Developing a Less Subjective Assessment. Transfund Report OG/0306/24S, Wellington, New Zealand.

usRAP (2009). usRAP Feasibility Assessment and Pilot Program, retrieved online at: http://www.usrap.us/home/.

Vogt, A. (1998). Accident Models for Two-Lane Rural Segments and Intersections. *Transportation Research Record 1635*, pp.18-29.

Wang, J., Sii, H.S., Yang, J.B., Pillay, A., Yu, D,, Liu, J., Maistralis, E., and Saajedi, A. (2004) . Use of Advances in Technology for Maritime Risk Assessment. *Risk Analysis* 24(4), pp. 1041–1063.

WSDOT (2009). *State Highway Log : Planning Report*. Retrieved online at http://www.wsdot.wa.gov/mapsdata/tdo/PDF_and_ZIP_Files/HwyLog2009.pdf.

WSDOT (2011). *WSDOT GeoData Distribution Catalog.* Accessed online at http://www.wsdot.wa.gov/mapsdata/geodatacatalog/

WHO (2010). World Report on Road Traffic Injury Prevention, retrieved online at http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/en/index.html.

Wu, H., and Zhang, Z. (2010). Impact of Delivering Road Risk Indices to Road Users on Traffic Safety: Simulation-Based Evaluation. Transportation Research Board 89th Annual Meeting, Washington, D.C.

Wu, H., Gao, L., and Zhang, Z. (2011). Analysis of Crash Data Using Quantile Regression for Counts. Presented in the 90th Transportation Research Board Annual Meeting, Washington, D.C., January, 2011.

Wu, Y., Wang, Y., and Levy, A. B. (2008). Accident Risk Modeling for Two-Lane Rural Roads in Washington State. *Transportation Research Board 87th Annual Meeting,* Washington DC.

Yannis, G., Kanellopoulou, A., Aggeloussi,K., and Tsamboulas,D. (2005). Modelling Driver Choices towards Accident Risk Reduction. *Safety Science* 43(3), pp. 173 – 186.

Zegeer, C.V., Stewart, J.R., Huang, H.H., and Lagerwey, P.A. (2001). Safety Effects of Marked vs. Unmarked Crosswalks at Uncontrolled Locations: Analysis of Pedestrian Crashes in 30 Cities (with discussion and closure). *Transportation Research Record 1773*, pp.56-68.

Zhang, L., and Levinson, D. (2008). Determinants of Route Choice and Value of Traveler Information: A Field Experiment. *Transportation Research Record 2086*, pp.81-92.